PETR TŘÍSKA

GENETIC LEGACY OF TRANS-ATLANTIC SLAVE TRADE IN PRESENT POPULATIONS: ANTHROPOLOGICAL AND CLINICAL CONTEXT

Tese de Candidatura ao grau de Doutor em Ciências Biomédicas submetida ao Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto.

Orientador – Doutora Luísa Pereira

Categoria – Investigadora

Afiliação – Instituto de Investigação e Inovação em Saúde, Universidade do Porto (i3S); Instituto de Patologia e Imunologia Molecular da Universidade do Porto (Ipatimup).

Coorientador – Doutor Pedro Soares

Categoria – Investigador

Afiliação – Instituto de Patologia e Imunologia Molecular da Universidade do Porto (Ipatimup); Department of Biology, CBMA (Centre of Molecular and Environmental Biology), University of Minho, Braga, Portugal.

Coorientadora – Professora Doutora Maria de Fátima Gärtner

Categoria – Professora Catedrática

Afiliação - Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto.

Research work coordinated by:

Financiamento:

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 290344 (EUROTAST).

EUROTAST

The History, Archaeology & New Genetics of the Transatlantic Slave Trade

Ao abrigo do ponto nº 2, alínea a) do Art.º 31.º do Decreto-Lei n.º 115/2013, de 7 de agosto de 2013, fazem parte integrante desta dissertação os seguintes manuscritos já publicados ou em preparação para publicação:

Triska P, Soares P, Patin E, Fernandes V, Cerny V, Pereira L. 2015. Extensive admixture and selective pressure across the Sahel Belt. Genome Biol. Evol. 7:3484–3495.

Sierra B, **Triska P**, Soares P, Garcia G, Perez AB, Aguirre E, Oliveira M, Cavadas B, Regnault B, Alvarez M, Ruiz D, Samuels DC, Sakuntabhai A, Pereira L, Guzman MG. *OSBPL10, RXRA* and lipid metabolism confer African-ancestry protection against dengue fever in admixed Cuban population. In preparation.

Fernandes V**, Triska P**, Pereira JB, Alshamali F, Rito T, Machado A, Fajkošová Z, Cavadas B, Černý V, Soares P, Richards MB, Pereira L. 2015. Genetic stratigraphy of key demographic events in Arabia. PLoS One 10:e0118625.

According to the relevant national legislation, the author declares that he actively participated in the execution, analyses and discussion of results, as well as in the elaboration of the publications, under the name Triska P. The author clarifies that he took the leading role in Paper 1, and wrote that paper. In paper 2, the author shared the leading role with co-author Sierra B (MD, PhD), having been responsible for the bioinformatic analyses in which the paper relays, and he participated in the writing of the manuscript together with Sierra B and Pereira L. In Paper 3, the author was responsible for the bioinformatic analyses of the genome-wide data, collaborating in the redaction of the paper in the relevant parts related with the autosomal information, but the final responsibility of writing the paper belonged to Fernandes V (PhD), Richards MB (PhD) and Pereira L (PhD). All co-authors revised and approved the papers that will only be included in this thesis. The remaining sections of this thesis were written by the author.

Hope is definitely not the same thing as optimism. It is not the conviction that something will turn out well, but the certainty that something makes sense, regardless of how it turns out.

Václav Havel

Acknowledgments

I would like to thank to all people who helped me during the work on research contained in this thesis.

Firstly, to my supervisors, Dr Luisa Pereira and Dr Pedro Soares for their patience, scientific guidance and supervision, which they provided me during almost four years of work on my PhD. Also to Professora Maria de Fátima Gärtner for help with ICBAS administration and useful comments to my work, to Professor Eduardo Rocha for the possibility of obtaining PhD in Biomedical Sciences, and to Professor Manuel Sobrinho Simões for opportunity to work in excellent environment of IPATIMUP.

I would like to thank to all members of our research group, as well as many other people from IPATIMUP, who gave me warm welcome when I arrived to Portugal and made me feel like home. I am especially thankful to students and post-docs, who helped me directly with my research and without them I would hardly accomplish my goals: Bruno, Veronica, Marisa, Orlando, Andreia and Joana.

Also, I would like to thank to all PIs and PhD fellows from EUROTAST project, who became a good friends of mine. Together we travelled to amazing places and the memories we share made my PhD years awesome.

Last but not least, I would like to acknowledge European Commission and Marie Curie Actions for funding my research. Without their financial support this research would not be possible.

Contents

List of Figures

Introduction

Paper I

Paper II

Paper III

Final Discussion

Supplementary Material Paper I

Supplementary Material Paper II

Supplementary Material Paper III

List of Tables

Introduction

Paper II

Paper III

Supplementary Material Paper I

Supplementary Material Paper II

Abbreviations

A

A – adenine

ABCA1 – ATP-binding cassette transporter ABCA1 gene

AD – *Anno Domine*

APOL1 – apoliporotein 1

ATP – adenosine triphosphate

B

BSP – Bayesian skyline plots

Bp – base pair

C

C – cytosine

CACNG3 – Voltage-dependent calcium channel gamma-3 subunit gene

CATSPERG – Catsper channel auxiliary subunit gamma gene

CCL3L1- Chemokine (C-C motif) ligand 3-like 1 gene

CD234 – Cluster of Differentiation 234 gene

CDK – Chronic Kidney Disease

CEU – Utah Residents (CEPH) with Northern and Western European Ancestry

CFTR – Cystic fibrosis transmembrane conductance regulator

CLM – Colombians from Medellin, Colombia

CMK1D – Type II Ca(2+)/calmodulin-dependent protein kinase D gene

CNV – copy-number variation

COL5A1 – Collagen, Type V, Alpha 1 gene

CV – cross-validation

CXCR4 – C-X-C chemokine receptor type 4 gene

D

DARC – Duffy antigen/chemokine receptor (DARC) gene

DENV – Dengue virus

DF – Dengue fever

DGAT2 – Diglyceride acyltransferase (or O-acyltransferase) gene

DGKI – Diacylglycerol kinase iota gene

DHF – Dengue haemorrhagic fever

DNA – deoxyribonucleic acid

DOCK10 – Dedicator of cytokinesis 10 gene

DRC – Democratic Republic of Congo

DSS – Dengue shock syndrome

E

ESN – Esan in Nigeria

EVX1 – Even-skipped homeobox 1 gene

F

FCG – fever comparison group

G

G – guanine

GA – Guantanamo asymptomatic
GH – Guantanamo haemorrhagic
GF – Guantanamo fever

GBR – British in England and Scotland

GC – Guantanamo control

GSEA – Gene Set Enrichment Analysis

GW – genome-wide

GWAS – genome-wide association study

GWD – Gambian in Western Divisions in the Gambia

H

HA – Havana asymptomatic

HBB – Hemoglobin beta gene

HbS – Hemoglobin sickle

HC – Havana control

HCG – haemorrhagic comparison group

HCV – Hepatitis C Virus

HDL – high density lipoproteins

HF – Havana fever

HGDP – Human Genome Diversity Project

HH – Havana haemorrhagic

HIV – Human Immunodeficiency Virus

HLA – Human Leukocyte Antigen

HOXA – Homeobox A Cluster gene

HVS-I – hypervariable region 1

HVS-II – hypervariable region 2

HWE – Hardy-Weinberg equilibrium

I

IDW – Inverse distance weighted

IFN – interferon gene

IgM – Immunoglobulin M

iHS – integrated haplotype score

INSIG2 – Insulin induced gene 2 gene

IRF3 – Interferon regulatory factor 3 gene

ITGAE – Integrin, alpha E gene

K

kb – kilobase

KCNJ12 – Potassium channel, inwardly rectifying subfamily J, member 12 gene

KEGG – Kyoto Encyclopedia of Genes and Genomes

KRT39 – Keratin 39, type I gene

KRT40 – Keratin 40, type I gene

ka – thousand years ago

L

LCT – lactase gene

LD – linkage disequilibrium

LDL – low density lipoproteins

LGM – last glacial maximum

lncRNA – long non-coding RNA

LWK – Luhya in Webuye, Kenya

M

MAF – minor allele frequency

MAPKAPK5 – MAP kinase-activated protein kinase 5 gene

Mb – megabase

MCM6 – Minichromosome Maintenance Complex Component 6 gene

MCMC – Markov Chain Monte Carlo

MHC – Main Histocompatibility Complex

MICA – MHC Class I Polypeptide-Related Sequence A gene

MICB – MHC Class I Polypeptide-Related Sequence B gene

mmHg – milimetres of mercury

mRNA – messenger RNA

MSL – Mende in Sierra Leone

MSY – Male-specific region of chromosome Y

mtDNA – mitochondrial DNA

MXL – Mexican Ancestry from Los Angeles USA

N

NCAPG2 – Non-SMC Condensin II Complex, Subunit G2 gene

NF-kB – Nuclear factor kappa-light-chain-enhancer of activated B cells

NPR3 – Natriuretic peptide receptor 3 gene

O

OAS – 2'-5'-oligoadenylate synthetase gene

OCG – overall comparison group

OR – odds ratio

ORF – open reading frame

OSBPL10 – Oxysterol binding protein-like 10 gene

P

PC – principal component

PCA – principal component analysis

PI3P – Phosphatidylinositol 3-phosphate gene

PIK3AP1- Phosphoinositide-3-Kinase Adaptor Protein 1 gene

PIK3R1 – Phosphoinositide-3-Kinase, Regulatory Subunit 1 (Alpha) gene

PLC – Phospholipase C

PLEKHG1 – Pleckstrin homology domain containing, family G gene

PLEKHM1L – Pleckstrin homology domain containing, family M, member 3 gene

PPNB – Pre-pottery Neolithic B

PRDM9 – PR domain zinc finger protein 9 is a protein gene

PTPRN2 – Receptor-type tyrosine-protein phosphatase N2 gene

Q

Q-Q – quantile-quantile

R

RAB3GAP1 – Rab3 GTPase-activating protein catalytic subunit gene

RHG – rainforest hunter-gatherer

RNA – ribonucleic acid

RSPO3 – R-spondin-3 gene

RT-PCR – real-time PCR

RXRA – Retinoid X receptor alpha gene

RYR2 – Ryanodine receptor 2 gene

S

SLC45A2 – Solute carrier family 45 member 2 gene

SLC24A5 – Solute carrier family 24 member 5 gene

SNP – single nucleotide polymorphism

SOX6 – Transcription factor SOX-6 gene

SPATS2 – Spermatogenesis associated, serine-rich 2 gene

SPINT2 – Serine peptidase inhibitor, Kunitz type, 2 gene

SREBF2 – Sterol regulatory element binding factor 2 gene

SRY – Sex-determining region Y protein

STAT – Signal transducer and activator of transcription

SYN3 – Synapsin III gene

T

T – tymine

TAS2R – Taste Receptor, Type 2 gene

TAS2R16 – Taste Receptor, Type 2, Member 16 gene

TAST – Trans-Atlantic Slave Trade

TMRCA – time to most recent common ancestor

TNF – tumor necrosis factor gene
TRPV1 – Transient Receptor Potential Cation Channel, Subfamily V, Member 1 gene
TSI – Toscani in Italia
U
U – uracil
UAE – United Arab Emirates
ULK4 – Unc-51 Like Kinase 4 gene
UV – ultra violet
V
VDR – Vitamin D receptor gene
VIPR2 – Vasoactive intestinal peptide receptor 2 gene
W
WDR60 – WD repeat domain 60 gene
WHO – World Health Organization
X
XPEHH – cross-population extended haplotype homozygosity
Y
YRI – Yoruba in Ibadan, Nigeria
Z
ZRANB3 - Zinc finger, RAN-binding domain containing 3 gene

Abstract

The African continent harbors a high level of human genetic diversity. The African genetic landscape was shaped by numerous migrations, admixture events and selective adaptations to environmental factors and pathogen burden. In Africa, the Sahel belt was one of the most important migration routes, while the Trans-Atlantic Slave (TAST) Trade introduced the largest forced migration of Africans into Europe and Americas, and the Arabic Slave Trade was a parallel phenomenon into Arabian Peninsula and the Near East. In this study we aimed to investigate the admixture patterns along these migration routes and the role that African ancestry played in the adaptation to environment and diseases.

We performed dense (2.5 million) SNP genotyping in 161 Africans from Sahel (west, central and east; sedentary and nomadic), and in 273 Cuban individuals in the context of dengue fever disease (controls, asymptomatic and dengue fever patients with and without hemorrhages). In addition, we sequenced 57 new whole mitochondrial DNA (mtDNA) samples from East Africa and Arabian Peninsula.

SNP data were analyzed for population stratification, linkage disequilibrium patterns, signals of positive selection and local admixture inference. Mitochondrial data were used for phylogenetic reconstruction and time to most recent common ancestor (TMRCA) analysis. Admixture analysis in Sahel indicated uniform African ancestry of Western groups and admixture with non-African ancestry in Central and Eastern groups. Local ancestry mapping in East Africans identified excess of African ancestry at *DARC* gene, providing complete resistance to malarial agent *Plasmodium vivax,* and a peak of non-African ancestry in *RAB3GAP1/LCT/MCM6* region possibly related with lipid metabolism. We also applied iHS and XP-EHH selection tests to investigate selection signals from haplotype structure. We have noted that several genomic regions exhibited a signature of positive selection in all investigated populations, while other signals are geographically clustered. In particular, strong selection signals on *DARC* and *PIGG* genes were observed in populations across the Sahel. On the contrary, selection signature on genes involved in lipid metabolism was confined to Eastern Sahel populations, while signal in *SPINT2/CATSPERG* region was specific for populations of Western Sahel. We also report a

population specific signal of selection on *TAS2R* taste receptors in Fulani possibly driven by sexual selection.

In Cuba, we identified two regions of African ancestry significantly linked with the asymptomatic dengue fever phenotype: a short haplotype in 3p22.3, within *OSBPL10* gene; and 9q34.3 region in proximity of the *RXRA* gene. For both genes, the most significant SNPs are placed outside the coding region, being probably involved in the gene expression regulation. Indeed, we confirmed that the expression of these genes changes along dengue disease progression, by measuring the mRNA expression of *OSBPL10* and *RXRA* in Cuban patients. The mRNA expression of *RXRA* was significantly lower along the disease than in convalescence (day 30). While mRNA expression of *OSBPL10* is significantly lower at day 3 but rises at day 7, and decreases again in convalescence. This evidence confirms the African protection against dengue hemorrhagic fever, which might be mediated through cholesterol and retinoid acid metabolism, essential for instance in the replication of viruses in hepatocytes and production of cytokines in macrophages.

Finally, we estimated migration events between Africa and Eurasia through information from mitochondrial and genome-wide data. We argue that founder analysis based on mitochondrial markers provide better estimates for old demographic events than methods based on linkage decay between segregating autosomal markers. The capacity to disentangling between recent and old migration events is essential to properly ascertain the impact of the African slave trade into several parts of the world, except into the Americas. Our analyses indicated that gene flow across the Red Sea was mainly due to the Arab slave trade and maritime voyages in the period from 2.5 ka to very recent times, but had already begun by the early Holocene.

The detailed ancestry characterization in African and African descendant populations is of high anthropological and medical relevance and can help us to dissect complex host-pathogen interactions and to shed light on several complex diseases.

Resumo

O continente africano comporta uma elevada diversidade genética humana. A paisagem genética africana foi moldada por numerosas migrações, eventos de mistura e adaptações seletivas a fatores ambientais e patogénicos. O Sahel constituiu uma das rotas mais importantes de migração dentro de África, enquanto o Tráfico Trans-Atlântico de Escravos (TAST) introduziu a maior migração forçada de africanos na Europa e nas Américas e o Tráfico Árabe de Escravos foi um fenómeno paralelo para a Península Arábica e o Próximo Oriente. Neste estudo tivemos como objectivo investigar os padrões de mistura ao longo destas rotas de migração e o papel que a ancestralidade africana desempenhou na adaptação a ambientes e doenças.

Assim, realizámos a genotipagem de alta resolução de SNPs (2,5 milhões) em 161 africanos do Sahel (oeste, centro e leste; sedentários e nómadas) e em 273 Cubanos no contexto da febre da dengue (controlos, assintomáticos e doentes com febre da dengue sem e com hemorragias). Adicionalmente, sequenciámos 57 genomas mitocondriais completos (mtDNA) em amostras do Leste de África e da Península Arábica.

Os dados dos SNPs foram analisados quanto a estratrificação populacional, padrões de *linkage disequilibrium*, sinais de seleção positiva e inferência de mistura local. Os dados de mitocondrial foram usados para reconstrução filogenética e estimativa do ancestral comum mais recente (TMRCA). A análise de mistura no Sahel indicou uma ancestralidade africana uniforme nos grupos oeste e mistura com ancestralidade não-africana nos grupos central e leste. O mapeamento por ancestralidade local nos africanos do leste identificou excesso de ancestralidade africana no gene *DARC*, que confere resistência ao agente da malária *Plasmodium vivax,* e um pico de ancestralidade não-africana na região *RAB3GAP1/LCT/MCM6*, possivelmente relacionada com o metabolismo de lípidos. Também aplicámos testes de seleção iHS e XPEHH de modo a investigar sinais de seleção a partir da estrutura dos haplótipos. Notamos que várias regiões exibem sinais de seleção positiva em todas as populações investigadas, enquanto outros sinais estão geograficamente agrupados. Concretamente, sinais fortes de seleção foram observados para os genes *DARC* e *PIGG* em todas as populações do Sahel. Em oposição, sinais de seleção em genes envolvidos no

metabolismo dos lípidos estavam confinados às populações do Leste do Sahel, enquanto sinais na região *SPINT2/CATSPERG* eram específicos do Oeste do Sahel. Também reportamos um sinal numa população, a seleção nos receptores de sabor *TAS2R* nos Fulani, possivelmente derivado de seleção sexual.

Em Cuba, identificamos duas regiões de ancestralidade africana associada com o fenótipo assintomático da febre da dengue: um haplótipo pequeno em 3p22.3, no gene *OSBPL10*; e na região 9q34.3, na proximidade do gene *RXRA*. Em ambos os genes, os SNPs mais significativos estão localizados fora da região codificante, estando provavelmente involvidos na regulação da expressão dos genes. De fato, através da medição da expressão de mRNA de *OSBPL10* e *RXRA* em doentes cubanos, confirmamos que a expressão destes genes se altera ao longo da progressão da doença da dengue. A expressão do *RXRA* era significativamente inferior durante a doença quando comparada com a convalescença (dia 30). Enquanto a expressão de OSBPL10 era significativamente inferior no dia 3, aumentava no dia 7 e voltava a diminuir na convalescência. Esta evidência confirma a proteção africana contra a febre hemorrágica da dengue, possivelmente mediada através do metabolismo de colesterol e ácido retinóico, essenciais por exemplo para a replicação dos vírus nos hepatócitos e a produção de citoquinas nos macrófagos.

Finalmente, estimamos es eventos de migração entre África e Eurasia através de informação mitocondrial e autossómica. Argumentamos que a análise de fundador baseada em marcadores mitocondriais providencia melhores estimativas para eventos demográficos antigos do que os métodos baseados em *linkage decay* para os marcadores autossómicos. A capacidade de discernir entre eventos migratórios recentes e antigos é essencial para avaliar corretamente o impacto do tráfico de escravos africanos para várias partes do mundo, exceto para as Américas. A nossa análise indica que o fluxo génico através do Mar Vermelho foi maioritariamente devido ao tráfico Árabe de escravos e às viagens marítimas no período desde 2.500 anos até quase ao presente, mas tinha já começado no Holoceno Inicial.

A caracterização detalhada das populações africanas e das suas populações descendentes é de elevada relevância antropológica e médica, podendo ajudar a dissetar interações complexas hospedeiro-patogénio e a esclarecer o mecanismo de várias doenças complexas.

1 Introduction

1.1 Human genome

1.1.1 Deoxyribonucleic acid (DNA) and the genetic code

Genetic information is stored in the molecule of deoxyribonucleic acid (DNA), composed of two complementary strands of biopolymers coiled around each other. Each strand is build up from nucleotides, the monomeric subunits consisting of nitrogen-containing nucleobase, deoxyribose and phosphate group. The phosphate group is connected to the sugar by covalent bond forming the alternating sugar-phosphate backbone of the DNA strand. Each deoxyribose sugar has a nitrogen-containing base attached to it. There are four types of nucleotides: adenine, guanine, cytosine and thymine. Nucleotides are grouped by the type of nitrogen-containing base: adenine and guanine are purines; cytosine and thymine are pyrimidines. The different chemical properties of purines and pyrimidines determine the pairing fashion between the nucleotides: purine adenine (A) pairs with pyrimidine thymine (T), and purine guanine (G) pairs with pyrimidine cytosine (C). Nucleotides in double-stranded DNA are connected by hydrogen bonds. The double stranded DNA is wrapped around structural proteins histones, forming organized structures of packed DNA called chromosomes.

During the process of transcription one strand of DNA is used as a template for synthesis of messenger ribonucleic acid (mRNA), which serves as a mediator of the genetic information between DNA strand and ribosome, where nucleotide sequence is translated into protein. The mRNA is a single-chain nucleic acid composed of the same nucleotides as DNA, except for thymine, which is substituted by uracil (U). The synthesis of proteins at the ribosomes is based on the translation of nucleotide sequence in mRNA to the sequence of amino acids of a protein. This process is defined by the genetic code that determines the correspondence between the 64 possible triplets and the 20 amino acids. Triplet or codon is a sequence of three consecutive nucleotides and each of them codes for a particular amino acid, except for the stop codons, which terminate the synthesis of protein. Since the number of amino acids is lower than the number of possible codons, the genetic code is redundant, i.e. one amino acid can be coded by several codons.

The human genome contains approximately 3.2 billion base pairs (bp) distributed into 23 chromosomes, located in the nucleus of cells (nuclear DNA; nDNA). In somatic cells, the human genome is present in diploid state, meaning there are two sets of genetic information and all chromosomes, present in pairs. The human karyotype comprises 22 pairs of autosomal chromosomes present in both sexes, and one pair of sex-specific chromosomes, whose

combination determines the sex. The female karyotype contains two chromosomes X, the male karyotype contains one X and Y chromosome.

The human genome harbours approximately 20-25 thousands of genes, the genomic regions coding for functional RNA or protein product. Genes constitute only approximately 1.5% of the human genome, while the remaining genome consists of non-coding RNAs, regulatory sequences, retrotransposons, introns, and sequences with unknown function.

The majority of the nDNA can be subjected to recombination during meiosis, by which portions of the paired chromosomes (one inherited from the father and the other from the mother) are shuffled, originating new combinations of genetic diversity. But there are portions of the genome that are inherited only from one parent and therefore escape the recombination process. These uniparental markers are the mitochondrial DNA (mtDNA) typically inherited only in the maternal lineage, and the Y chromosome inherited exclusively in the paternal lineage.

1.1.2 Mitochondrial DNA

The mitochondrion is a double-membrane cell organelle located in the cytoplasm of nearly all eukaryotic cells. Human cells contain approximately 100 mitochondria, involved in several cellular processes. Primarily, they supply the cell with chemical energy stored in the form of adenosine triphosphate (ATP), however mitochondria are also involved in signaling, cellular differentiation and cell cycle (Giles et al. 1980). Each mitochondrion contains 5-15 copies of mtDNA, which is organized in a circular double stranded molecule that contains 16,569 bp, coding for 37 genes. Out of those genes, 13 are polypeptides of the oxidative phosphorylation system, 22 are transfer RNAs and 2 ribosomal genes. The genetic code used for the translation of mtDNA sequence into proteins differs from the code used for nDNA (Ernster & Schatz 1981). In particular, triplets AGA and AGG are stop codons, UGA encodes for tryptophan, and AUA codes for methionine (Chinnery et al. 1999). Also, the largest portion of genes required for mitochondrial function (more than 90%) is located in the nuclear genome, including genes involved in the oxidative phosphorylation (Gray 1993).

Because mtDNA does not undergo recombination and rapidly accumulates mutations (the mtDNA mutation rate is approximately 10 times higher than the average rate of nuclear genome (Jobling et al. 2013)), it is frequently used as a marker in population genetics (Torroni et al. 2006). It allows for inferences about the past migrations and demographic events through statistical comparison of variation between mtDNA sequences. Particularly important is the

coalescence estimation, which is based on the assumption that all human mtDNA sequences descended from a single ancestral known as mitochondrial Eve, retroactively reconstructing evolution of sequences back in time. Additionally, given the roughly constant mutation rate of the mitochondrial genome along human evolution, coalescence algorithms can provide estimation of the time elapsed since a group of sequences shared a common ancestor (TMRCA). Individual branches and sub-branches of human mtDNA sequences are classified into haplogroups, forming the human mtDNA tree, which reflect the history of human migrations. Most basal haplogroups L0, L1 and L2 are confined to Africa. Haplogroup L3 is mostly African as well, but it also encapsulates all of the non-African haplogroups, included in sub-haplogroups M and N. This fact led to the model of out of Africa migration (Cann et al. 1987), which is vastly accepted nowadays: the origin of modern humans took place in Africa, and the colonization of the remaining globe occurred after migration from Africa. Haplogroup M is frequently found throughout South Asia, Siberia and Northeast Asia. Subgroup M1 is present also in East Africa, plausibly because of back migration from Eurasia (Richards et al. 2006).

Haplogroup N contains major sub-haplogroup R, which covers the majority of European mtDNA lineages, and non-R haplogroups, which constitute the deepest branches in haplogroup N, as the rare haplogroups N1 and N2, present for example in isolated Siberian populations, and haplogroups O and S found in Australasia.

1.1.3 Y chromosome

Y chromosome is a sex chromosome specific for males. Its length is around 58 million bp, 90% of which compose the non-recombining region. It contains more than 200 genes, 72 out of them having known protein product. Particularly important is the sex determining region (SRY) gene, which initiates male sex determination during embryonal development.

Analogically to mtDNA, Y chromosome phylogeny reflects the out of Africa migration and colonization of continents. The Y chromosome haplogroups are labeled from A to T, where A is the most basal group and T is the most divergent. Therefore, the deepest split in the human Y chromosome tree is between the haplogroup A, which is restricted to African and African descendant populations, and groups from B to T, which can be found also outside the African populations. Haplogroup A has the highest frequency in Namibian San and Nama people with frequencies 66% and 64%, respectively (Wood et al. 2005). High frequencies of A haplogroup were reported also from Sudanese Dinka, Shilluk and Nuba people (Hassan et al. 2008).

Haplogroup B is also predominantly African with high frequency in hunter-gatherer populations Biaka and Mbuti (Berniell-Lee et al. 2009), but was reported also from Arabian peninsula (Abu-Amero et al. 2009) and Eurasia (Grugni et al. 2012). In Middle East, haplogroup J is frequent, which can be found also throughout South Europe. In general, the most frequent haplogroups in Europe are I and R, although R is also common in Central Asia and Indo-European speakers from North India. Haplogroup N is often found in Siberian and Northeast Asian populations like Even, Nenets and Buryat. The Native Americans have high frequency of haplogroups Q, and in Northern America lineages from haplogroup C are also present, although they are very frequent in Central and North Siberians, Australian Aborigenes, New Zealand Maori and Polynesians. Southeast Asia is dominated by lineages belonging to haplogroup O, except New Guinea, where the most prevalent is haplogroup M (Jobling et al. 2013).

1.2 Processes shaping genetic diversity of populations

1.2.1 Mutation

The only source of novel genetic variation is mutation, any change in the genetic information producing a new allele (Jobling et al. 2013). Mutation occurs as result of the failure in the DNA replication or DNA repair mechanisms. According to the scale of the DNA alteration, mutations can be classified into several categories.

1.2.1.1 Point mutation

The most frequent type of point mutation consists in the exchange of one single nucleotide by another (also single nucleotide polymorphism or single nucleotide variant, SNP or SNV). Depending on the type of nitrogenous base, point mutations can be either transitions or transversions, the first much more common (Freese 1959). In transition, a purine nucleotide is substituted by another purine, or a pyrimidine by another pyrimidine. On the contrary, transversion is a substitution of a purine by a pyrimidine or vice versa. Although there are two possible transversions and only one possible transition for each nucleotide, transitions are more frequent, because substituted nucleotides have equal number of ring structures. Nucleotide substitutions can be caused either by misincorporation of the nucleotide during DNA replication or by chemical/physical mutagens. Misincorporation can occur when DNA polymerase attaches the wrong nucleotide at the 3' daughter strand and the proofreading mechanism of the replication machinery fails to recognize this misincorporation. The high fidelity of the replication mechanism ensures that replication errors occur less than 1×10^{-9} per nucleotide per replication event (McCulloch & Kunkel 2008) in the human genome.

Besides misincorporation during replication, point mutations can result also from spontaneous endogenous chemical processes in cells, which can trigger the alteration or loss of the nucleotide base. This includes deamination, oxidation, methylation and depurination. Chemical mutagens include substances which can be incorporated during the replication but having different base-pairing properties (base analogs), substances altering base-pairing properties (base modifying agents), molecules which can insert between nucleotides and distort helix structure (intercalating agents), or substances which cross-link different parts of the DNA helix (cross-linking agents) (Jobling et al. 2013).

Apart from chemical mutagens, the DNA is exposed also to a variety of physical influences with mutagenous potential, in particular low-energy electromagnetic radiation (e.g. ultra violet radiation) and higher-energy ionizing radiation (e.g. gamma radiation). Adverse effects of UV

radiation arise from its ability to induce DNA lesions through cyclobutane–pyrimidine dimers (Sinha and Häder 2002). Ionizing radiation can promote mutations directly by inducing DNA strand breaks or indirectly by producing free radicals, which in contact with the DNA molecule can cause nucleotide substitutions or deletions.

Insertions or deletions of one base can also occur, and are designated by indels.

1.2.1.1.1 Synonymous mutations

Single nucleotide polymorphisms and indels might affect the function of genes. There is a wide variety of possible results of point mutations, depending whether the mutation leads to changing the sequence of amino acids and, consequently, the final protein. In the case of synonymous mutations, the sequence of amino acids remains intact, due to the already mentioned high redundancy of the genetic code. This mechanism provides a certain degree of protection against the adverse effects of base substitutions. Although synonymous mutations are often considered silent, under certain circumstances they can affect transcription, splicing and translation, and in this way alter the resulting phenotype (Chamary et al. 2006).

1.2.1.1.2 Nonsynonymous mutations

Point mutations which alter the sequence of amino acids are called nonsynonymous or missense. If the replaced amino acid has similar chemical properties (e.g. hydrophobicity), the mutation is considered conservative. But even a conservative substitution can have a considerable functional impact if it occurs at binding sites of the protein.

A non-conservative substitution occurs when one amino acid is substituted by an amino acid with different chemical properties. This kind of mutation is more probable to be deleterious than the conservative substitutions, as the resulting protein is likely to lose its function.

The most severe type of point mutation in terms of effect on protein product is indel in an open reading frame (ORF). Insertion or deletion of one or more bases in non-multiples of three leads to shifts in the reading frame, hence the final protein of mutated phenotype can be completely different from the wild type protein, depending on the position of mutation in gene (worse at the beginning of the ORF). Frame shifts can also lead to premature termination of transcription, if a stop codon is reached, or on the contrary, to abnormally large protein if the stop codon is omitted (Jobling et al. 2013). If the deletion is in multiples of three, it will lead to the loss of one or several amino acids, but to no frame shift of the protein.

Among typical examples of deletions within ORF leading to severe disease is cystic fibrosis (Dalemans et al. 1991), which is mainly caused by deletion of phenylalanine amino acid

(ΔF508) from the CFTR protein, leading to protein malfunction. Another example is the deletion of 6 bp in *APOL1* gene, resulting in two neighbouring amino acids being removed from apolipoprotein L-1, which leads to higher risk of kidney disease, but also provides resistance to sleeping sickness (Genovese et al. 2010).

1.2.1.2 Structural variation

Genomic changes larger than 1 Kbp are usually regarded as genomic structural variations. Based on the scale and nature of the genomic alteration, structural variation can be classified into several categories:

1.2.1.2.1 Microscopic structural variation

This type includes large genomic changes that can be observed by optical microscope, such as aneuploidies and chromosomal aberrations (Reich et al. 2002). Chromosome aberration (abnormality) is the most severe type of genomic structural variation. Typically it involves deletions or duplications of whole parts of chromosomes, often with pathologic effect on the phenotype. Aneuploidity indicates a state, where one or more chromosomes are present in a different count than two. For instance, trisomy in chromosome 21 leads to Down syndrome, or monosomy of X chromosome determines Turner syndrome.

1.2.1.2.2 Copy-number variation

Copy-number variation (CNV) is a type of structural variation, where organism have abnormal number of copies of certain portions of DNA. Investigation of CNVs is complex and requires specific cytogenetic techniques e.g. fluorescent in-situ hybridization or comparative genomic hybridization (Sudmant et al. 2010). Although most of the CNVs probably do not affect phenotype, some of them have been associated with susceptibility or resistance to certain diseases. It has been shown, that individuals with higher number of copies of *CCL3L1* gene have decreased risk of infection and/or progression of HIV infection (Dolan et al. 2007). Around 5% of the human genome consists of duplicated DNA sequences enriched for genes participating in immunity response and some of them have dosage effect (Gonzalez et al. 2005).

1.2.2 Meiotic recombination

During meiotic recombination, homologous chromosomes can either exchange portions of chromosome in the process of crossing-over, or convert a part of sister chromatide in non-reciprocal manner in the process of gene conversion. Recombination between homologous chromatides breaks down haplotypes in recombination locations and creates new recombined

haplotypes on both homologous chromosomes. Recombination is more likely to occur between distant loci on chromosome and less likely between proximate sites, but the recombination rate is not uniform along the chromosome, creating patterns of recombination hotspots and zones of low recombination (**Figure 1**). Analysis of fine-scale recombination map revealed a particular DNA sequence of 13 bp, which has been associated with recombination hot-spots in humans: CCNCCNTNNCCNC. This motif is possibly a binding site for protein PRDM9, which is capable of modifying histone H3 and, in this way, trigger recombination (Myers et al. 2008).

Thus, meiotic recombination plays an important role in evolution, with sexual reproduction producing high levels of diversity in every generation and allowing faster adaptation than in asexually reproducing organisms (Jobling et al. 2013).

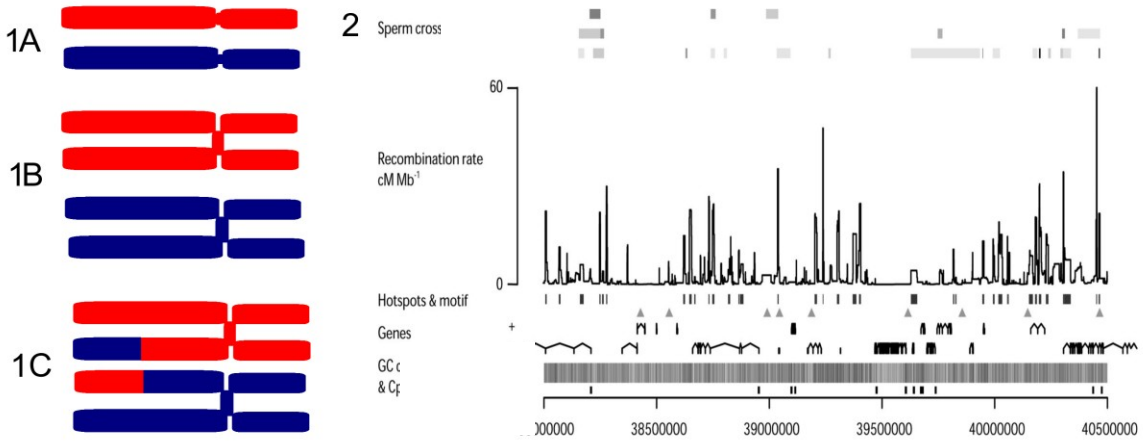

Figure 1 - Chromosomal recombination by crossing over. (1A) Chromosomes in interphase G1. (1B) Chromosomes in interphase S. (1C) Chromosomes during crossing-over in prophase 1. (2) Variability of recombination rate along chromosome 21, with indication of hotspots (adapted from Myers et al. 2006).

1.2.3 Linkage disequilibrium

As the recombination between nearby loci is less probable than between distant loci, neighbouring alleles are linked. This non-random association of alleles is called linkage disequilibrium. Investigation of linkage disequilibrium provides useful information in association studies, because due to linkage of nearby alleles, fewer typed SNPs are needed to flag associated loci. Linkage disequilibrium is also a good indicator of positive selection, because advantageous haplotypes can produce patterns of long-ranging linkage. The most common measures of linkage disequilibrium are Lewontin's D, D' and r^2. D is the simplest measure of linkage disequilibrium, defined as a difference between observed and expected frequency of

two-locus haplotype:

$$D = x_{11} - p_1 q_1$$

Since the value of D depends on allele frequency, comparison of LD between different loci requires standardization of values. This is done by D', where the absolute value of D is divided by the maximal value:

$$D' = |D| / D_{max}$$

Another measure is r^2, calculated as the squared correlation coefficient between two loci.

1.2.4 Genetic drift

In contrast to the latter mentioned processes which enhance diversity, genetic drift is an evolutionary process which acts in the opposite direction and decreases allele diversity in the population gene pool. Genetic drift is a change in allele frequency in population in time due to the random sampling of organisms in the next generation (Masel 2011). The strength of the genetic drift is determined by the effective size of the population: small and isolated populations are more likely to experience stronger genetic drift than the large ones (Zimmer 2001, **Figure 2**).

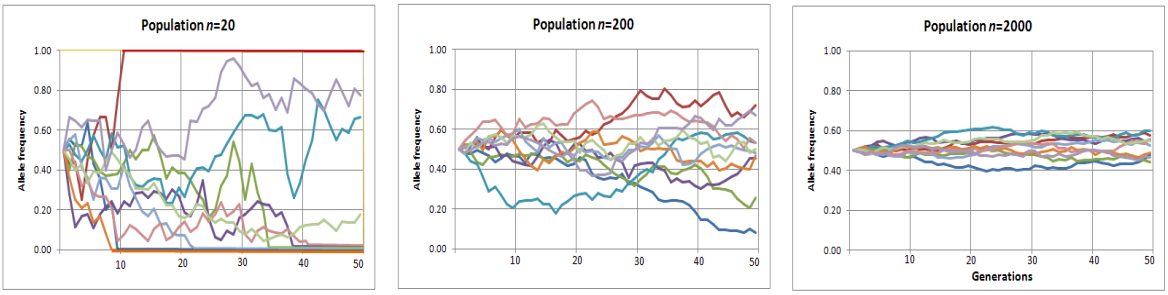

Figure 2 – Simulated genetic drift. Graphic representation of computer-simulated genetic drift in populations of 20, 200 and 2000 individuals over 50 generations. The effect of genetic drift is stronger in smaller than larger populations (CC BY-SA 3.0 via Commons license).

1.2.5 Bottleneck and founder effect

Bottleneck and founder effect are demographic events with strong impact on genetic diversity. Both are characterized by rapid decrease in population size. In the case of bottle neck, the population size decreases for example due to epidemics or abrupt environmental changes, while founder event results from colonization of a new habitat. In both cases, the genetic diversity of the resulting population is a subset of the genetic diversity present before the event.

1.2.6 Gene flow and migration

Migration is a movement of individuals or populations in space from one inhabited area into another (Jobling et al. 2013). If an individual contributes to the next generation in the new location, this event is called gene flow. In context of population genetics, several models are used to describe complex processes of migration and subsequent gene flow.

The simplest model describing gene flow is the n-island model introduced by Wright (1931). This model assumes that the metapopulation is divided into a number of sub-populations of equal size which exchange genes at equal rate. This model further assumes that none of the sub-populations can go extinct, there is no geographical sub-division apart from islands and that there are no evolutionary processes like mutation or selection. If assumptions of this model are met, the rate of migration *m* can be directly related to F_{ST}:

$$F_{ST} = \frac{1}{1 + 4\mathrm{Nm}}$$

The stepping-stone model is an advanced simulation of gene flow, since it accounts also for geographic distance. Similarly to the n-island model, it assumes an even rate of migration between the sub-populations, but it considers that the islands are arranged into a matrix and that the gene flow occurs only between neighbouring islands. This model was further developed into the isolation by distance (IBD) model, where the genetic similarity between populations is a function of dispersal distances (Wright 1943). The simulations based on models of gene flow proved that even a small amount of gene flow between populations can slow down their genetic differentiation (Wright 1950; Jobling et al. 2013).

1.2.7 Natural selection

Natural selection is an evolutionary process in which advantageous genes are more likely to be passed to the following generation (and eventually become fixed in the population) as a result of greater fitness of the bearer of that trait, while disadvantageous traits are less likely to be passed down to the next generation (and eventually become eliminated), since bearer of disadvantageous trait is less likely to mate successfully and bring up an offspring (Darwin 1872). This mechanism helps populations to get rid of deleterious mutations and to adapt to the new environmental conditions and pathogens.

Selection acts during all stages of life of the individual important to reproduction (based on Jobling et al. 2013): ability to survive into reproductive age; ability to attract sexual partner; ability to procreate; and number of offspring.

1.2.7.1 Purifying selection

The majority of mutations have deleterious effect and negatively influence function of proteins. This is reflected also in the fitness of individuals carrying mutated variant: strongly deleterious variants derogate viability of the individual, and the damaging mutation is eventually eliminated from the population. This type of selection reduces diversity around conserved regions.

1.2.7.2 Negative selection

Some of the standing variation may become disadvantageous under the new environmental conditions and consequently undergo negative selection. For example, carriers of blood type O are significantly more susceptible to the cholera infection than other blood types. As the cholera infection imposes strong selective pressure, the variant coding blood group O was negatively selected in the region of delta of river Ganges and nowadays is almost absent in the local population (Harris et al. 2008).

1.2.7.3 Balancing (overdominant) selection

Balancing selection favours high diversity in locus and acts against fixation of one allele. The resulting genomic signature is high diversity, measured as high proportion of alleles with intermediate frequency and paucity of low and high frequency alleles.

Certain mutations are harmful in homozygous state, but on the other hand increase fitness of heterozygous individual. This variant is under balancing selection, maintaining equilibrated frequencies of ancestral and derived allele in the population. One of the best studied examples is HbS variant in gene for haemoglobin B, where hydrophilic glutamic acid is replaced by hydrophobic valine. This leads to malformation of erythrocytes, which form sickles instead of normal round shaped cells. In homozygous state, this disease causes severe anaemia and significantly decreases life expectancy, but also confers resistance to malaria. In heterozygous state, the malaria resistance is preserved, while causing only mild anaemia, which, in terms of fitness, is outweighed by strong advantage of resistance to malaria (Kwiatkowski 2005). This convenience of heterozygous state maintains both alleles in the population as long as the evolutionary pressure (in this case malaria parasite) is present.

Another example of balancing selection is a region on chromosome 6 coding major histocompatibility complex (MHC). It is one of the key elements of the immune system of vertebrates. Molecules of MHC are responsible for binding of peptides extracted from pathogens and displaying them on the cell surface, where these can be recognized by T-cells (Janeway et al. 2001). Higher variation in MHC region is advantageous for an organism, as

this allows to recognize broader range of pathogens. Hence, the amount of diversity of MHC region in populations is often correlated with diversity of pathogens in the area where the population resides (Prugnolle et al. 2005).

1.2.7.4 Positive selection

Variants which enhance fitness of an organism undergo positive selection. This means that the variant is likely to gradually increase its frequency in the population.

1.2.7.4.1 Hard sweep

A selective sweep is the reduction of haplotype diversity around the region under the positive selection (Voight et al. 2006; Pritchard et al. 2010). In case of hard sweep, a new advantageous mutation arises and quickly becomes frequent in the population. The rapid increase of frequency prevents recombination from breaking down the haplotype around the selected allele. The resulting genomic signature is long-ranging haplotypes and extended haplotype homozygosity (**Figure 3**).

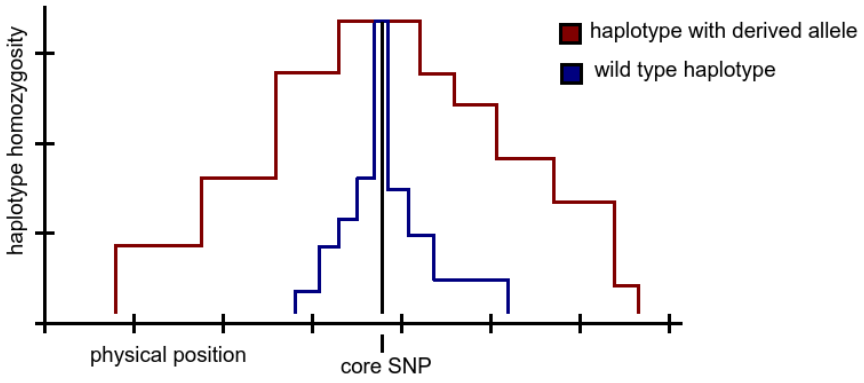

Figure 3 – Genomic signature of hard sweep. A new advantageous variant (red) rapidly becomes frequent in the population, while increasing frequency of carrying haplotype. This creates the genomic signature of long haplotypes and high homozygosity, compared to wild type (blue). Based on Karlsson et al. (2014).

1.2.7.4.2 Soft sweep

In the process of soft sweep, positive selection acts on variants which have been present in the population for some time, the so called standing variation. If there are several advantageous alleles, which confer roughly equivalent amount of selective advantage, none of these alleles will rapidly reach fixation, but the sum of frequencies of these alleles will increase over time (Messer & Petrov 2013).

1.2.7.4.3 Polygenic adaptation

Polygenic adaptation is the type of response to positive selection when several genes code for the selected trait and become selected simultaneously. As a result, all advantageous variants of the selected genes will gradually increase their frequency (Pritchard et al. 2010).

1.2.7.4.4 Genomic regions under recent positive selection

Within the last 100,000 years, human populations have repeatedly migrated into new environments, what required extensive adaptations to a wide range of different climates, diet and pathogens (Voight et al. 2006; Sabeti et al. 2007). Additionally, the onset of farming ~8 ka ago brought a radical change in lifestyle, as farmers had to face pathogens transmitted from domesticated animals and important changes in diet (Diamond 1997). The availability of next-generation sequencing data is allowing to investigate signals of positive selection in numerous modern and ancient populations at the genome-wide level, and pointing out candidate genes for selection (Pritchard et al. 2010; Mathieson et al. 2015). In this sub-chapter we present some of the well-known examples.

1.2.7.4.4.1 Lactase persistence (LCT)

Human ability to digest sugar present in fresh milk (lactose) is determined by the production of an enzyme lactase-phlorizin hydrolase, which splits the molecule of disaccharide lactose into two simple sugars: glucose and galactose (Vesa et al. 2000). This enzyme is normally expressed in brush cells in the small intestine by infants. After the weaning, the production of lactase enzyme is usually inhibited. However, some populations with tradition of pastoralism and fresh milk consumption have high proportion of individuals who retain an ability to produce the lactase enzyme also in adulthood (lactase persistence, Tishkoff et al. 2007). This gives them an evolutionary advantage over the lactase non-persistent individuals, since lactase persistent individuals can exploit fresh milk as a source of nutrients. Lactase persistence is prevalent in Europe (mainly Northern), Arabia, Central Asia and East and West Africa (**Figure 4**). It has been shown, that lactase persistence is associated with several point mutations upstream the *LCT* gene: G-22018 and T-13910 in European populations and C-14010, G-13907 and G-13915 in East Africa (Ranciaro et al. 2014). Extensive association studies in Africa and Europe revealed, that European and African variants determining lactase persistence lie on different haplotype backgrounds and result from convergent evolution (Tishkoff et al. 2007). *LCT* region was identified as one of the strongest signals of positive selection in the human genome: in populations where *LCT* was selected, variants involved in LCT persistence sit on unusually long and frequent haplotype (Voight et al. 2006; Sabeti et al.

2007).

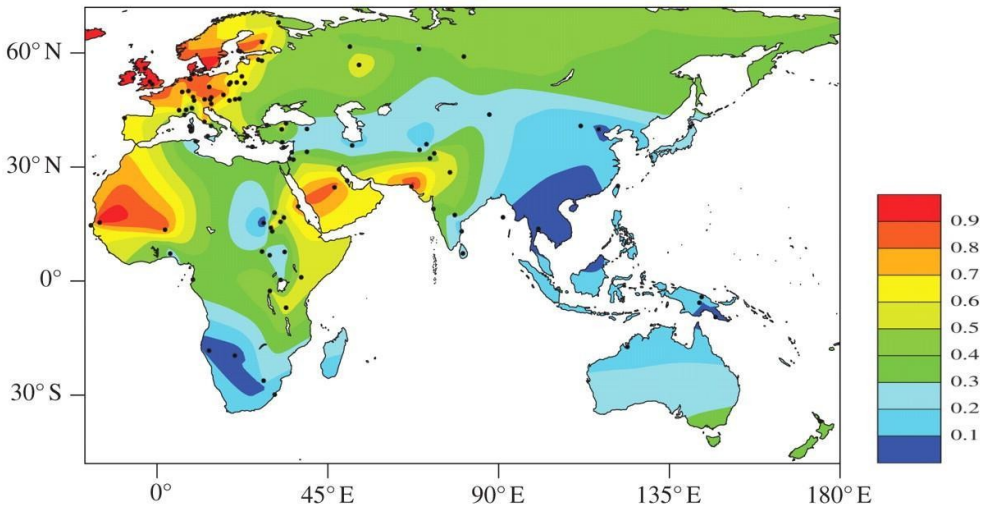

Figure 4 - Interpolated frequency of lactase persistence phenotype around the Old World. Image adapted from Gerbault et al. (2011).

1.2.7.4.4.2 Light skin color in Europeans

Skin color is determined by the density and size of melanosomes, vesicular objects filled with the pigment melanin, present in the upper layer of skin. Pigmentation of skin plays an important role in absorption of the high-energy UV rays, which cause harm to the different cellular structures, e.g. cause DNA damage and induce malignant mutations (Bernstein et al. 2002). The level of pigmentation in the human populations correlates with the amount of UV radiation in the area. It has been proposed that light skin color in higher latitudes facilitates the synthesis of vitamin D, which regulates calcium absorption and calcium levels in bones (Dawson-Hughes et al. 1995). In the European populations, skin color is largely determined by variation in two genes from the solute carrier family: *SLC24A5* and *SLC45A2*. *SLC24A5* mediates production of melanin through regulation of calcium in melanocytes, while *SLC45A2* plays a role in processing melanin precursors (Jablonski & Chaplin 2000). Both these genes have been under strong positive selection in Europe (Voight et al. 2006; Pritchard et al. 2010; Mathieson et al. 2015), as indicated by the following evidence: high F_{ST} scores, high derived allele frequency and extended haplotype homozygosity. The exact place and time of origin of the selected derived alleles in *SLC24A5* and *SLC45A2* is not clear. Analysis of ancient DNA samples revealed that West-European hunter-gatherers living between 7-8 ka had ancestral alleles in both genes and therefore dark skin. However, Swedish hunter-gatherers from the roughly same time period already had derived alleles, and presumably had light skin. In samples from the

Neolithic farming populations, the derived variant in *SLC24A5* has been identified as frequent (nearly fixed), but only low to intermediate frequencies of the derived variant in *SLC45A2* were detected (Mathieson et al. 2015).

1.2.7.4.4.3 Duffy antigen system

Duffy antigen chemokine receptor (*DARC*, *CD234*) is a non-specific receptor on the surface of erythrocytes. It is also the receptor for *Malaria vivax* parasite. Duffy system has three codominant alleles: FY*A, FY*B and FY^{ES}. The FY^{ES} allele denotes Individuals homozygous for FY^{ES} allele that do not express *DARC* receptor on the surface of erythrocytes and hence are resistant to infection by *Malaria vivax* (Jobling et al. 2013). This gives carriers of Duffy-null blood group strong advantage, and led to nearly-fixation of the FY^{ES} allele in Sub-Saharan Africa (Nickel et al. 1999; Howes et al. 2011, **Figure 5**). The disadvantage of Duffy-null phenotype is that carriers of this phenotype are more prone to be infected by HIV-1 virus (Kulkarni et al. 2009; Ramsuran et al. 2011).

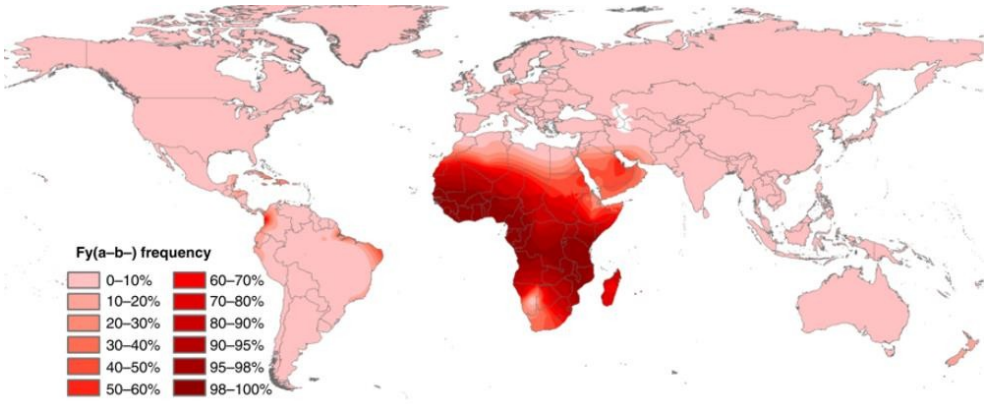

Figure 5 - Worldwide distribution of Duffy-negative phenotype. Besides sub-Saharan Africa, Duffy-negative phenotype is present in Arabian Peninsula and in some areas of South America. Adapted from Howes et al. (2011).

1.2.7.4.4.4 Bitter taste receptors

The ability to detect inedible or poisonous compounds in food is an important prerequisite of successful survival, especially in a new environment with previously unknown species of plants and animals. In mammals, this ability is mediated by olfactory and taste receptors, which evaluate chemical properties of ingested food. In this context it is particularly relevant the perception of bitter taste, a common attribute of various inedible plant substances such as glycosides or alkaloids (Li & Zhang 2014).

The bitter taste is mediated by seven transmembrane receptors, coded by 43 members of the *TAS2R* gene family on chromosomes 7 and 12. The *TAS2R* genes located on the same chromosome are often very similar, which indicates series of duplication events in the evolution of this gene family (Fischer et al. 2005; Bachmanov & Beauchamp 2007). The bitter taste receptors are expressed on taste buds in the oral cavity, but also in parts of the respiratory tract, suggesting a pleiotropic function of this gene family, probably linked to the innate immunity (Deshpande et al. 2010; Lee & Cohen 2014).

Scans for positive selection revealed strong signal on several members of *TAS2R* gene family in Africa and Eurasia, although the pattern of selection acting on the taste receptors is different between populations (Wang et al. 2004; Li & Zhang 2014). For example, *TAS2R16* gene coding for sensitivity to the bitter compound salicin has been shown to be under strong positive selection in East Africa (excess of derived allele), but under purifying selection in West and Central Africa (absence of non-synonymous mutations in ancestral haplotype). While in East African populations the derived variant 516-T which codes for ability to taste salicin is more common, in West and Central Africa, the ancestral variant 516-G which determines the non-tasting haplotype is more prevalent. The Fulani population from Cameroon is an interesting exception that, despite being a Central African population, exhibits strong signal of positive selection on *TAS2R16* and also the allele frequencies are similar to those in East African populations (Campbell et al. 2014).

1.3 Analysis of genetic diversity in genome-wide data

1.3.1 F_{ST} statistics

F_{ST} is a common statistics widely used as a measure of genetic diversity found within subpopulations compared to the total population (Jobling et al. 2013). F_{ST} ranges between 0 and 1 and can be inferred from the genetic diversity data by several methods. Usually F_{ST} is computed as:

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

where H_T is the expected heterozygosity of the whole population and H_S is the expected heterozygosity in the subpopulations. Pairwise F_{ST} can be also used as a measure of genetic distance, in this case it is defined as:

$$F_{ST} = \frac{V_p}{p(1-p)}$$

where p is the mean and Vp is the variance of the gene frequencies between the populations.

According to published guidelines, F_{ST} values in terms of genetic differentiation should be interpreted as follows: less than 0.05: little genetic differentiation; 0.05-0.15: moderate; 0.15-0.25 great; more than 0.25: very great (Wright 1965).

Pairwise F_{ST} can be visualized either as a Manhattan plot, showing F_{ST} values separately for each SNP, or as a heat map, showing mean F_{ST} between pairs of populations (**Figure 6**). The Manhattan plot representation is useful in the identification of variants highly divergent between populations, for example in selection studies, while the heat map gives us information on the overall genetic similarity between the studied populations

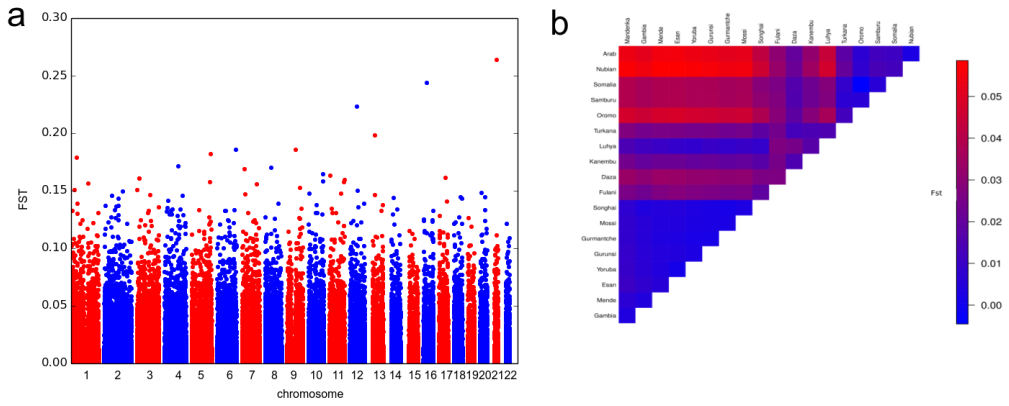

Figure 6 - Visual representations of F_{ST} values. a) Manhattan plot of F_{ST} values for each SNP; b) heat map of mean F_{ST} values between pairs of populations.

1.3.2 Principal component analysis

Principal component analysis (PCA) is a statistical method for reduction of multi-dimensional data into a certain number of vectors while preserving as much information as possible. The method is based on the identification of correlated vectors and in transforming them into principal components (eigenvectors), which capture the main trends in data. The weight of individual eigenvectors is expressed by the eigenvalue, which is proportional to the amount of variance explained by the respective eigenvector. For visualization of population structure, it is usually used the combination of the first three PCs: PC1 vs PC2 and PC1 vs PC3. Nevertheless, remaining eigenvectors may contain an important amount of information as well, and the distribution of eigenvalues needs to be considered.

1.3.3 Clustering methods STRUCTURE and ADMIXTURE

Model based clustering methods use the prior assumption of a defined number (K) of source populations with characteristic sets of allele frequencies. The ancestries of all individuals in the data set are then inferred as a composition of K modeled source populations. The model assumes no linkage disequilibrium (implying pruning the genome-wide datasets for LD) and complete Hardy-Weinberg equilibrium within populations. The algorithm then aims to assign every allele in every individual to its putative origin from the source populations. This specifies the probability distribution of the model:

$$Pr(X|\, Z, P)$$

where X is the observed genotype, Z is the putative origin of allele and P is the allele frequency

in assigned source population. Parameters Z and P are decided randomly at the beginning of computation and then updated after every iteration of the Markov Chain Monte Carlo (MCMC) until the algorithm reaches the convergence criteria (Pritchard et al. 2000; Alexander et al. 2009).

The accuracy of the model fit to the data can be decided by the value of cross-validation (CV) error. In the cross validation procedure, a subset of individuals is not used for inference of the model parameters and these unobserved samples are used for estimation of accuracy of the model parameters.

1.3.4 Admixture dating methods

Admixture between genetically divergent populations creates patterns of long-range admixture linkage disequilibrium in admixed populations. Due to meiotic recombination, the length of these ancestral tracts becomes shorter with every successive generation. The algorithm of ROLLOFF (Moorjani et al. 2011) examines the decay of correlation between difference in allele frequency and LD between source populations. The observed correlations are used for fitting the exponential distribution, which is solved as a function of time and provides the estimation of time in generations that elapsed since the admixture took place.

1.3.5 Genome wide association study

A genome wide association study (GWAS) aims to identify variants associated with a disease/phenotype based on the information retrieved from characterizing a high amount of SNPs distributed across the whole genome. The design of GWAS is based on the statistical comparison of genotypes between case and control groups. The case group is made by individuals carrying the studied trait (e.g. disease phenotype), while the control group includes individuals without the trait. In order to avoid spurious results, both groups must be precisely matched in terms of age, sex and ancestry background. The algorithm compares the difference of allele frequencies between cases and controls and assigns a p-value to every SNP. Because every SNP is tested independently, the power of GWAS studies is limited by the large testing burden. This needs to be addressed by appropriate correction of p-values, otherwise false positive associations could be reported. The most conservative is Bonferroni correction, where the corrected statistical significance corresponds to uncorrected statistical significance level (usually 0.05) divided by the number of independent tests (Jobling et al. 2013). As chips used currently in GWAS have thousands or millions of SNPs, the significance threshold is usually 5

x 10^{-8} (Clarke et al. 2011).

GWAS studies conducted on various complex diseases revealed that the effect of a single locus on the disease phenotype is usually weak and a large number of samples is required in order to obtain p-values reaching the genome-wide threshold. The power of the GWAS can be increased by meta-analysis of data sets from independent studies. However, this approach contains several setbacks, as this usually requires merging data typed on different genotyping platforms with different set of SNPs. In addition, sampled groups might have different geographic origin, different case and control screening for studied trait, etc. Poor overlap of typed variants between two studies is usually improved by imputation of missing SNPs by inference from haplotype structure of reference panel, typically HapMap (Hastie et al. 2012). However, imputation has limitations, especially when populations analyzed are not well characterized. This is especially important in African populations, for which the amount of complete genomes available poorly represents the high level of genetic diversity within this continent. The standard visualization of GWAS results is a Manhattan plot (**Figure 7a**), where every tested SNP represented by a point: the X coordinate represents the physical position in chromosome and the Y coordinate represents the p-value usually in -log(10) scale. The distribution of p-values is routinely plotted into Q-Q (quantile-quantile) plot, where the Y axis represents the observed p-values and the X axis represents the p-values expected under uniform distribution (**Figure 7b**). The region with highest p-values can be plotted in LocusZoom (Pruim et al. 2010), which allows the visualization of associated SNPs in context of genes and recombination rate surrounding the associated loci (**Figure 7c**). A weak point of GWAS studies is that the SNP chips contain only common variants and therefore fail to identify association on rare variants. Some phenotypic traits are known to be heritable, but variants found by GWAS can explain only fraction of its heritability. Although the reason for this “missing heritability” is unknown, it has been suggested that the explanation (among others) might involve rare variants not represented in SNP chips, DNA methylation and RNA editing (Eichler et al. 2010).

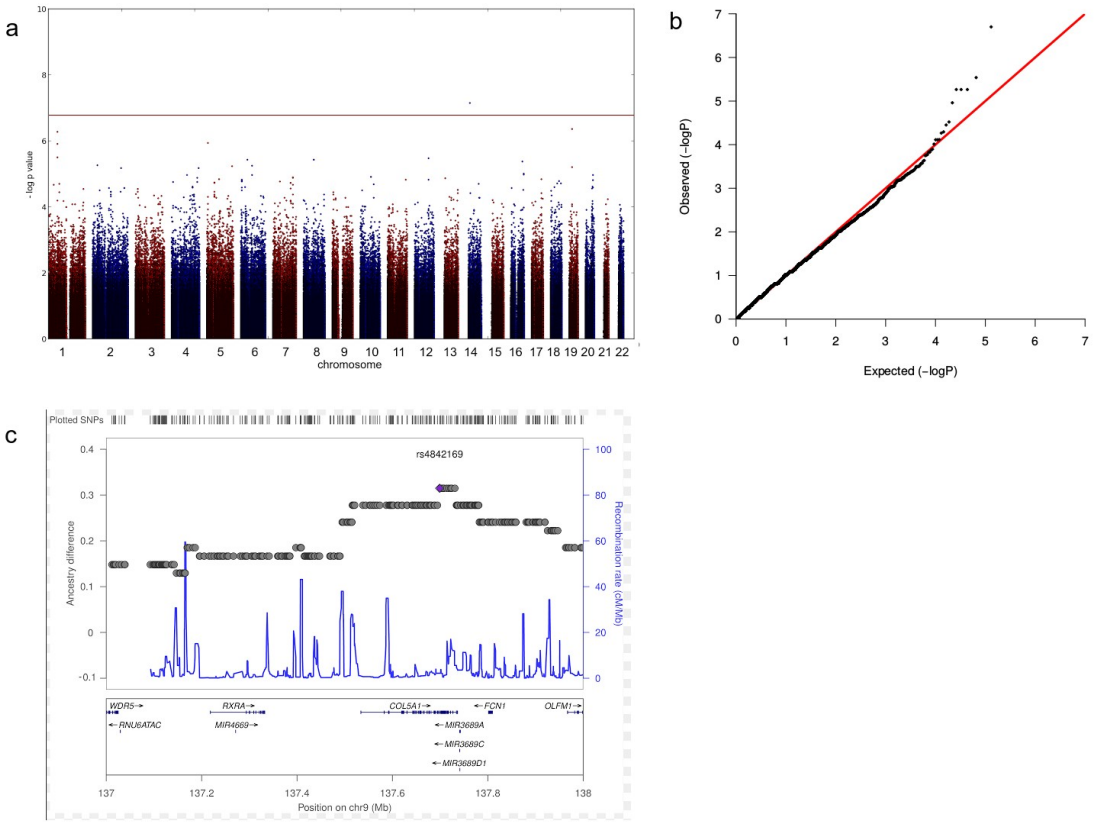

Figure 7 - Graphical representation of GWAS results. (a) Manhattan plot. (b) Quantile-quantile plot. (c) Locus Zoom. Images from Dengue study presented in this work.

1.3.6 Admixture mapping

Mapping of admixture LD is an approach frequently used in investigation of ancestry-related complex diseases in recently admixed populations, at the genome-wide level. The basic assumption of admixture mapping is that the genetic variant responsible for the investigated phenotype trait is population-specific and is surrounded by the admixture LD (Winkler et al. 2010). Admixture mapping statistically evaluates differences in local ancestry along the chromosome between case and control groups (**Figure 8**). An important advantage of admixture mapping is a considerably lower testing burden, because admixture blocks typically span across a considerable number of SNPs and therefore fewer tests are needed (Montana & Hoggart 2007).

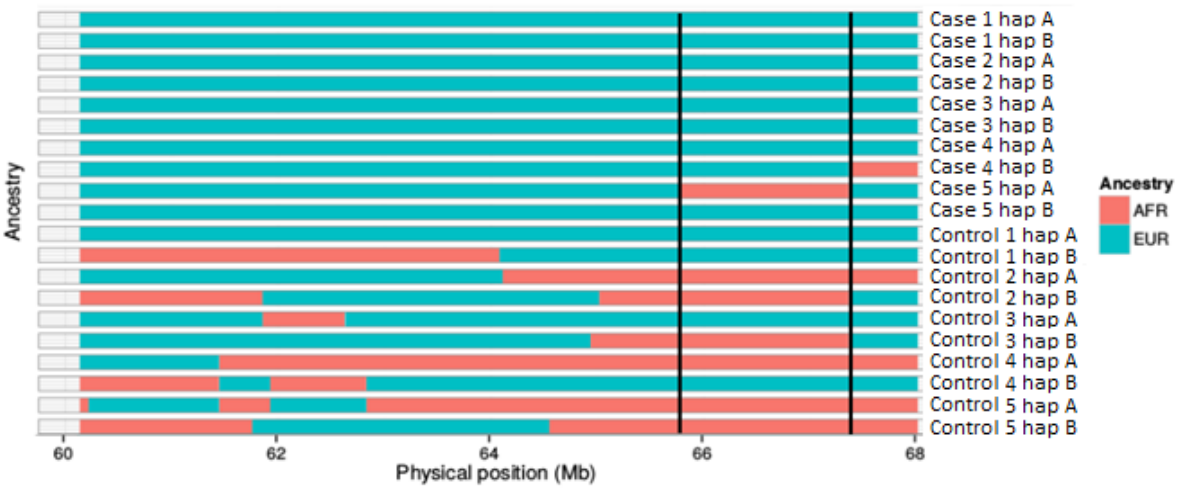

Figure 8 – Graphic representation of results of admixture mapping. Each horizontal bar represents portion of chromosome. Red color represents African ancestry tracts, blue represents European ancestry tracts. The candidate region (highlighted by vertical lines) will have a significantly higher amount of one ancestry (in this case, African) when comparing cases and controls.

Figure 5 - Language families in Africa. Adapted from "African language families en" by Mark Dingemanse. Licensed under CC BY 2.5 via Wikimedia Commons.**Figure 6 – Graphic representation of results of admixture mapping.** Each horizontal bar represents portion of chromosome. Red color represents African ancestry tracts, blue represents European ancestry tracts. The candidate region (highlighted by vertical lines) will have a significantly higher amount of one ancestry (in this case, African) when comparing cases and controls.

1.4 Worldwide population structure

1.4.1 Genetic structure of human populations

The history of the human species is reflected in the overall genetic diversity harboured in populations. The evolutionary approach enables us to reconstruct the early movements of human populations, even when the archaeological evidence is sparse or non-existent. This is possible thanks to the specific signature which various historical events (e.g. rapid change in population size or admixture with other populations) leave in the genetic structure of populations and which can be addressed by statistical approaches. Genetic structure between populations arises from accumulation of genetic variability from random mutations and decrease of gene flow between populations as geographic distance increases. Genetic structure is further shaped by evolutionary forces, in particular by natural selection and genetic drift. The biogeography of *Homo sapiens* is determined by several key demographic events and migrations, which founded the genetic structure of human populations as we know it today.

1.4.2 Out of Africa

Anatomically modern humans emerged ~150-200 ka in East Africa and for several tens of thousands of years, *Homo sapiens* lived in small groups (Newman 1995). Fossil evidence shows that anatomically modern humans were present in the Near East ~ 100-130 ka, but these early expansions did not persist. Studies of mitochondrial DNA suggest, that the first successful Out of Africa migration probably happened only 61-65 ka (Fernandes et al. 2012).

Anatomically and behaviorally modern humans reached Australia around 50 kya, Western Europe around 41 ka and China around 39 ka (Jobling et al. 2013). This expansion was facilitated by low sea level, especially in South-East Asia, where the large biogeographic region Sunda and Australia with Papua and Tasmania were connected in one landmass called Sahul (Soares et al. 2008). In fact, during the Pleistocene period the sea level was approximately 120 meters below current level as a result of climatic conditions of glacial maximum, during which huge portion of Earth's water was trapped in extensive ice caps (Ehlers & Gibbard, 2003).

America was the last habitable continent to be settled by humans. Some authors proposed that approximately around 18 kya, a group of migrating humans crossed Beringia straight from North-East Asia and began colonization of Americas (Goebel et al. 2008). Recent whole-genome studies demonstrated that native Americans descend from at least three colonization waves from Asia (Reich et al. 2012), although the exact dating of the migration waves remains

uncertain.

1.4.3 Population structure of Africa

Population structure of Africa is the most complex out of all continents. This is caused mainly by the fact that modern human species evolved in Africa and also accumulated the major part of the diversity observed worldwide.

1.4.3.1 Archaic hunter-gatherers

Most archaic populations living in Africa are groups of hunters and gatherers, sparsely distributed in regions of Central, East and South Africa.

Central Africa is home to populations of rainforest hunter-gatherers (RHG), often referred as “pygmies”. Beside the forest-dwelling subsistence, RHG populations are characterized also by typical short-stature phenotype: the average height of adult male is below 150 cm. This extreme phenotype is probably caused by several underlying genetic factors, in particular by the low expression of growth hormone receptors (Bozzola et al. 2009). It has been suggested that the short stature might be an evolutionary adaptation for the environment of dense tropical forest, and this hypothesis is supported by studies which observed patterns of selection on height-related genes in RHGs (Mendizabal et al. 2012; Migliano 2013). Populations of RHG are subdivided into two main branches: East and West RHG. The East branch of RHG comprises (among others) Mbuti, Batwa and Bakinga, who reside in Democratic Republic of Congo (DRC) and Uganda. Members of Western RHG populations are Bakola (Cameroon), Baka (Cameroon) and Biaka (Central African Republic) (Patin et al. 2014). The West and East branches of RHGs have been separated for a long time, around 20-30 ka (Patin et al. 2009).

South African San and East African Hadza and Sandawe speak click languages. Mitochondrial and Y chromosome lineages found in these groups occupy basal position in the phylogenetic tree of uniparental markers, indicating that click-language speakers are older than any other existing human population (Semino et al. 2002). Despite the millennia-long gene flow between hunter-gatherer populations and agriculturalists, they retained a unique genetic composition, distinct to all other African populations (Henn et al. 2011). Although anthropological and linguistic links between these populations is dubious, the genome-wide studies confirmed that these populations are genetically related (Pickrell et al. 2012).

1.4.3.2 Bantu speakers

The major population cluster in Africa belongs to Bantu-speakers (**Figure 9**). Bantu is a sub-branch of Niger-Congo languages, widely spoken in all regions of Africa south of equator. The

expansion of Bantu speakers to the East and South Africa began approximately 5 kya. From the source between Cameroon and Nigeria, two streams of Bantu expansion originated: one followed the western African coast southwards, while the other headed eastwards, towards the region of the Great Lakes and then southwards along the eastern African coast (Plaza et al. 2004; Tishkoff et al. 2009). The Bantu expansion changed dramatically the population structure of Africa by replacing important portions of diversity present in Central and South Africa, although some admixture with autochthonous populations occurred along the wave of dispersion. For example, the East African population Luhya is a Bantu speaking population which migrated to East Africa and admixed with East African populations, while maintaining Bantu language, and to a large extent, also Bantu (West African) genetic composition (Gurdasani et al. 2015).

1.4.3.3 Eurasian influence in East and North Africa

The population structure of East Africa has been heavily influenced by extensive contacts with Near East. Eurasian ancestry has been identified in almost all of the East-African populations studied so far. Although the exact source of Eurasian ancestry in East Africa is unclear, the gene flow had been probably quite extensive, as Eurasian ancestry reaches up to 50% in some Ethiopian populations (Pickrell et al. 2014). Back-to-Africa migrations followed by admixture of African and non-African populations probably happened during several independent events: first as early as the Last Glacial Maximum (LGM; Hodgson et al. 2014; our results presented in this work) followed by the Neolithic expansion from the Near East (our results; Arredi et al. (2004)). Early admixture has been confirmed also by a sequence of an ancient genome from an Ethiopian man who lived approximately 4,500 years ago: the genome of this individual contained substantial amount of West Eurasian ancestry, similar to early Neolithic farmers (Llorente et al. 2015).

The North African region has been populated during successive migration waves (Henn et al. 2012). As documented from the maternal gene pool, the initial settlement occurred approximately 40 ka during the Back-to-Africa migration from the Near East (Olivieri et al. 2006), followed by West Eurasian lineages that entered North Africa roughly at the end of LGM at 14 ka (Harich et al. 2010) and Near East populations that reached North Africa during the Neolithic (Arredi et al. 2004). These Back-to-Africa migrations gave rise to autochthonous North African populations like Berber in Morocco or Mozabite in Algeria. The gene pool of North Africa was later reshuffled by Arabs, who invaded North Africa in 7 th century AD and held the rule

over the territories until the rise of European colonialism. Arabs maintained vivid trade connections with West Africa, which included also slave trade, through trans-Saharan routes. Thus, West African mitochondrial lineages are still present in North Africa as a result of sex-biased mating, typical for slave trading in other regions of the globe (Henn et al. 2012). Shortly after the Arab expansion, Islam was adopted by autochthonous Saharan populations and later spread to West and Central Africa.

1.4.3.4 Sahel migration corridor

Sahel is an ecoclimatic zone in the transition between Sahara and savannas. It was established around 4,000 years ago, when climatic change in Africa fostered desertification of the Sahara. It spans from the Red Sea to the Atlantic coast, with a total length around 5,400 km. The average altitude of Sahel is between 200 and 400 meters with restricted mountain ranges. The relatively flat topography and the less extreme climate compared to Sahara made Sahel a convenient migration corridor in Africa.

Two different subsistence systems co-exist in the Sahel region: sedentary farmers and nomadic pastoralists. The earliest archaeological evidence of pastoralist subsistence in Africa comes from Central and Eastern Sahara (Wendorf et al. 1984; Mohammed-ali & Khabir 2003). From there nomadic pastoralists expanded to Sahel during Middle to Late Holocene. The nomadic life style of pastoralists results from climatic conditions of Sahel, characterized by long and dry periods interrupted by short rainy seasons. During the wet season, the southern fringe of Sahara becomes a suitable pasture for grazing the cattle. After the short rainy season, herders must seek for pastures in wetter regions in the southern part of Sahel (Bučková et al. 2013). Sahel is one of the few regions, where the nomadic pastoralist subsistence is practiced until present day. The largest pastoralist group in Sahel is Fulani (also called Fulbe or Peul) people. The Fulani population probably originated from a Saharan population around 5 kya. In the 11th century AD, groups of Fulani settled and founded the kingdom Tekrur at Senegal river (Cerný et al. 2006). Fulani speak Fulbe language, which belongs to the Atlantic branch of Niger-Congo languages. Although originally nomadic, many of them are nowadays sedentary and live in urban areas. Despite being one of the biggest ethnic groups in West Africa, Fulani are a minority in all West African countries. The genetic composition of Fulani is derived from West African background similar to Wolof and Mandinka from Senegal, with an important proportion of Eurasian ancestry (Gurdasani et al. 2015). Eurasian genetic ties are apparent also in the

maternal gene pool of Fulani, where West-Eurasian haplogroups U5, H, J1b and V were reported (Cerný et al. 2006).

More nomadic pastoralist groups are found in central and eastern part of Sahel. In northern Chad around the Lakes of Ounianga live the semi-nomadic herders Daza (Podgorná et al. 2013), in northern Kenya around the Lake Turkana live the semi-nomadic pastoralists Samburu and Turkana.

As indicated by paleobotanical records, farming is in Sahel region more recent than pastoralism (Neumann 2003). Farmer populations are dispersed along southern and wetter parts of Sahel and exchange products with pastoralists, although intermarriage is not common and gene flow is limited (Cerný et al. 2011).

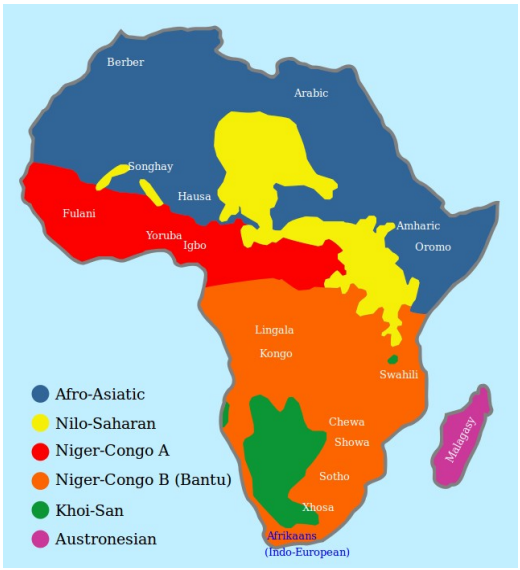

Figure 9 - Language families in Africa. Adapted from "African language families en" by Mark Dingemanse. Licensed under CC BY 2.5 via Wikimedia Commons.

1.4.4 Colonization of Europe

The demographic history of Europe depended largely on climatic conditions on the continent after the Out of Africa expansion. Pleistocene climate was characterized by recurrent cycles of cold glacial stages and warm interglacials with periods of about 100 kya. While warm interglacial periods were relatively short and typically more temperate, during the glacial stages the ice sheets advanced, sea levels and precipitation dropped and average temperatures were considerably lower than today (**Figure 10**)

After the warm interstadial event ~110-130 kya, the temperatures were declining and reached the minimum values during LGM 23-14 kya, when sea levels were ~ 120 below present level and most of North Europe was covered by extensive ice sheets. As most of the ice-free Europe was covered by polar desert, there were only few locations hospitable for temperate biota. These areas known as glacial refugia bear distinguished milder (micro) climatic characteristics such as higher precipitation and higher temperatures. The most important glacial refugia were located in Mediterranean peninsulas (Iberia, Italia and Balkans), but there is also a good amount of evidence for presence of smaller refugia in south-facing valleys of Alps (Rodríguez et al. 2009) and Carpathians (Sommer and Nadachowski 2006), valley of river Rhone (Michaux et al. 2004) and Crimean peninsula (Sommer and Benecke 2005).

Genetic evidence suggests that East Asian population and West Eurasian diverged ~45-36.2 ka (Fu et al. 2014; Seguin-Orlando et al. 2014). Archaeological records document human presence in Europe as early as 40 ka (Jobling et al. 2013). The European Paleolithic population was limited by resources produced by hunting and gathering and the climatic conditions, which were not favourable until the end of LGM. This population only started to grow at the end of LGM when humans expanded from glacial refugia and began to colonize northern latitudes of Europe. Clinal pattern in Y chromosome haplogroups R1b1b2 and R1a1 observed in European modern populations is linked to post-glacial expansion from Iberia and Ukraine, respectively (Wei et al. 2013).

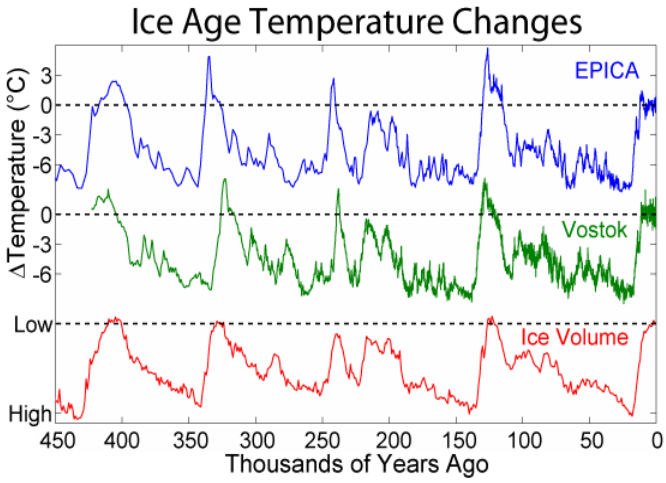

Figure 10 – Fluctuations of ice volume and surface temperature during the Pleistocene glacial cycles. Image based on Petit et al. 1999. Original image: "Ice Age Temperature". Licensed under CC BY-SA 3.0 via Wikimedia Commons.

Nevertheless, ancient DNA data from several Neolithic burial sites suggest, that there is discontinuity in Y chromosome lineages between Neolithic and modern populations, and that the spread of R1a1b2 haplogroup can be more recent (Jobling et al. 2013). Evidence from mitochondrial DNA points to repopulation from the Iberian refugium (Pereira et al. 2005) and the Near East (Pala et al. 2012).

About 10,200 years ago Near Eastern populations started the practice of farming subsistence system. This major cultural novelty was initially supposed to be followed by a large migration of farmers, and around 9000 years ago the first farmers from the Near East arrived in Europe and started its Neolithic transformation (Barbujani et al. 1998). From Greece, the migration of first farmers seems to have followed two streams: one into southeast Europe and second along the Mediterranean coast. The expansion continued for approximately 3,000 years, until Neolithic farmers reached British islands around 6,000 years ago (Jobling et al. 2013).

Around 4500 years ago a massive wave of steppe herders of Yamanaya culture migrated westwards from the Caspian region, replacing Neolithic populations present in Central Europe. Yamanaya people admixed with Central and North Europeans and led to formation of Bronze Age cultures like Corded Ware culture, Únětice culture. The fact that these cultures exhibit negative f3 admixture values supports a direct genetic link between Yamanaya people and latter West and Central European Bronze Age cultures (Allentoft et al. 2015). On the contrary, influence of steppe expansion had very limited or no effect in South Europe. For example, modern Sardinians appear to be genetically similar to Neolithic farmers from Italy, without any evidence of steppe ancestry (Allentoft et al. 2015). Thus, current European gene pool consists of at least three different ancestries: West European hunter-gatherers, Near Eastern Neolithic farmers and ancient North Eurasians. Proportions of two of these ancestries show clinal pattern: ancestry of early European farmers is strongest in Mediterranean reaching 90% and weakest in Baltic (30%). Western European hunter-gatherer ancestry is not present in Near East, however the ancient North Eurasian ancestry reaches up to 29% in Caucasus (Lazaridis et al. 2014). Yamnaya component is strongest in Northwest Europe and declines with latitude, while Neolithic component increases, reaching maximum at Sardinia (Haak et al. 2015). The resulting population structure in Europe therefore mirrors the geographic distribution of populations, where neighbouring populations are genetically related and genetic distance increases more along north-south axis compared to east-west axis (Novembre et al. 2008).

1.4.5 Asia

Human migration into Asia started shortly after the Out of Africa migration. Modern humans moved along the coast of Indian Ocean and through Southeast Asia, which was colonized approximately around 50-55 ka. By the 40 ka, the fully modern humans spread to most of the Old World including Europe, Central and East Asia and South Siberia.

Migration of modern humans through Southeast Asia was facilitated by the fact that this region was organized into a landmass called Sunda because of the low sea level during the glaciation cycle. When the temperature on Earth increased at the end of Last Glacial Maximum, around 15 ka, the sea level started to rise and turned Sunda landmass into the numerous archipelagos of Island Southeast Asia (Hewitt 2000). This flooding of landmass later triggered large-scale migrations in the region, leading to the Austronesian expansion which completely changed population structure in Southeast Asia (Soares et al. 2008).

1.4.6 Americas

1.4.6.1 South America

The displacement of approximately 11 million Africans to American colonies together with centuries-long immigration of European settlers shaped the demography of the New World after 1492. After abolition of slavery, populations in colonies gradually became mixtures of indigenous Native American populations, Europeans and descendants of enslaved Africans. Contributions of African and Native American men and women to the genetic pool of the colonial populations were not equal. Excess of maternal and lack of paternal African and Native American lineages points to strong mating biases in American colonial populations (Batista dos Santos et al. 1999). For example, in Brazil, the social policy between 17th and 20th centuries encouraged the marriage of Afro-Brazilian and Native American women with European men (Mörner 1967), while the low social status of men of Native American and African descent limited their access to women and marriage. The disproportion between paternal and maternal lineages of African and Native American origin in Brazilian population has been repeatedly documented (**Table 1**), but the same pattern of mating bias was observed also in Colombia (Carvajal-Carmona et al. 2000, Bedoya et al. 2006) Argentina (Salas et al. 2008), Cuba (Mendizabal et al. 2008) and Venezuela (Bortolini et al. 1999).

Table 1 - Ancestry of uniparental markers in Brazilian population.

European	African	Native American	region	marker	reference
57.9%	25.3%	16.8%	Rio de Janeiro	mtDNA	Bernardo et al. 2014
38.9%	27.9%	33.2%	Brazil, all regions	mtDNA	Alves-Silva et al. 2000
27.6%	35.2%	36.9%	Southeastern Brazil	mtDNA	Fridman et al. 2014
97.5%	2.5%	0%	Brazil, all regions	Y chromosome	Carvalho-Silva et al. 2001
88.1%	7.9%	4%	Rio de Janeiro	Y chromosome	Silva et al. 2006

Since the ancestry inferred from uniparental markers is highly biased in Latin America, the ancestral proportions inferred from autosomal genome-wide SNP data provide a better picture of the genetic structure of the populations. Analysis of genomic ancestry across different Brazilian states revealed homogenous pattern in the distribution of ancestries: European ancestry dominated, ranging between 60.6% and 77.7%; African ancestry accounted for 12.7%-30.3%; and Native American for 9.1%-19.4% (Pena et al. 2011).

In former Spanish colonies of Latin America that were not so heavily influenced by slave trade, like Mexico, Colombia and Ecuador, the African component is generally low (less than 10%) and populations are typically composed of Native American and European ancestries, where European ancestry on average accounts for ~40-60% and Native American for ~30-50% of total autosomal ancestry (Bryc et al. 2010b). Outside this pattern, stands locations like Mendellin in Colombia and Central Valley in Costa Rica, where high amounts of European ancestry (more than 60% on average) were documented, and on the contrary, Salta in Argentina or Lima in Peru with more than 60% of Native American ancestry (Wang et al. 2008, 1000 Genomes Consortium 2015).Even higher levels of Native American ancestry ranging up to 86% were documented in Bolivians from La Paz (Heinz et al. 2013) and isolated populations of Native Americans like Xavante, Suruí or Karitiana, which still have very low levels of genetic admixture (Kuhn et al. 2012; Reich et al. 2012).

1.4.6.2 Caribbean

Because the majority of the enslaved Africans were disembarked in the Caribbean, this region has also the highest amounts of African ancestry in America, although the distribution is highly

heterogeneous: while Puerto Rico and some regions of Cuba have only around 20% of African ancestry, in Haiti, Jamaica and Saint Thomas, the African ancestry ranges between 80 and 90% (Benn-Torres et al. 2008; Wang et al. 2008). Native American ancestry in the Caribbean islands is low, because the indigenous population of the islands was exterminated soon after the colonization.

Besides the general effect of three-population admixture pattern throughout all the Latin America, Trans-Atlantic slave trade (TAST) gave rise to several very unique populations of African descendants, who escaped from their slave masters and founded free communities of Marrón, usually in inaccessible areas like mountains or deep tropical forest. Although Marrón often intermarried with local Native Americans and adapted some kind of creole language, a mixture of European and African languages, Marrón were able to preserve their original African culture, including religion and traditions (Diouf 2014).

The first Marrón communities originated in Jamaica in the 17 th century, and later on also in Haiti, Cuba, Puerto Rico, Suriname and Brazil. Only a few Marrón communities survived to present day, mostly located in French Guyana and Suriname, together accounting approximately to 50 000 people. These communities preserved high levels of African genetic ancestry, both on maternal and paternal side (Brucato et al. 2010). In addition, African ancestry of Marrón people can be traced also to the African strains of human specific viruses Marrón people harbour (Martel-Jantin et al. 2014).

1.4.6.3 USA

According to historic records more than 305 thousands of Africans were disembarked in British North America, present day United States, mainly during the course of 18th and 19th centuries (Eltis & Richardson 2010). Communities of enslaved Africans worked primarily on sugar, tobacco and cotton plantations in southern states until December 1865, when slavery was abolished throughout the United States. Nowadays, approximately 42 million self-reported African Americans live in the United States (Rastogi et al. 2011). Their genetic ancestry is on average 78% West African, 19% European and 3% Native American, although there is large variation within individuals (Bryc et al. 2010a). The highest amounts of African ancestry were observed in African Americans from southeast states of Florida, South and North Carolina, Georgia and Alabama. Some of the African Americans have also traceable amounts of Native American ancestry, in particular in Oklahoma, where many Native American populations were relocated in the 19th century (Bryc et al. 2014).

Similarly to the other populations with history of slave trade, African Americans also exhibit

asymmetrical contributions of African males and females to their gene pool: European ancestry on Y chromosome was found in 28.46% of African Americans, but European mtDNA haplogroups were found only in 8.51% of examined individuals (Lind et al. 2007).

1.5 History of the Trans-Atlantic Slave Trade

1.5.1 Maritime discoveries

Between 14th and 15th century European navigators started to expand from coastal territories into the Atlantic Ocean. In the year 1402 Castilians began to colonize the Canary Islands, in 1419 Portuguese discovered Madeira and in 1427 Azores Islands. Vivid trade activity in North Africa created the need for new naval trading routes to West Africa in order to bypassing Moorish traders. Furthermore, expansion of the Ottoman Empire in the beginning of the 15th century disrupted the established trade routes between Asia and Europe and brought the need for alternative trade routes. Demand for new maritime routes fostered the development of nautical technology and navigation techniques. Advance in construction of ships led to the development of small and agile vessels called *caravelas* in mid-15th century, which were capable of windward sail and were better suited for an open ocean than any other type of ship at that time. This type of vessel was designed in Portugal, based on fishermen ship. The *caravela* ships were equipped with two or three masts with lateen sails, and the average length ranged between 12 to 18 meters.

In 1444 Portuguese reached Senegal and founded a small settlement on Goreé Island, near the present day city of Dakar. Around 1456 Cape Verde islands were discovered and in 1471 Portuguese navigators reached Elmina (**Figure 11**) in the Gold Coast (present day Ghana). Eleven years later, Diogo de Azambuja founded fortress São Jorge da Mina at this site, which soon become a center of trade for gold, ivory and slaves at the African Coast.

During the last decade of the 15th century, Spanish made important maritime discoveries in West Indies. In order to avoid conflict between Spanish and Portuguese, in 1494 both maritime powers signed a treaty, where Portuguese and Spanish divided discovered territories between them, known as the Treaty of Tordesillas. In this treaty, Spanish were given control over the western part of Americas and Pacific islands, while Portuguese were granted Brazil, Africa and Asia. This division of the world led these two potencies to establish TAST, and just in the first half century, between 1451 and 1499, it brought as many as 60,000 slaves to Europe (Rawley & Behrendt 2005).

Figure 11 - Elmina fortress on Ghanian Coast. This structure served as a main headquarters for Slave Traders in Golden Coast. Photo by Petr Triska.

1.5.2 First Atlantic system

In the first phase of TAST e almost exclusively Spanish and Portuguese traders were involved. In 1549 Portuguese started to establish sugar plantations in tropical and semitropical regions of Brazil. Sugar production created a massive demand for labour force. The first attempts to saturate the need for labour force with the Native Americans failed, both in Spanish and Portuguese colonies. Native Americans fought fiercely against European colonizers, and if enslaved, often rebelled or ran away from their masters. Furthermore, Native Americans had low levels of immunity against diseases brought by Europeans, such as flu and smallpox (Klein 1999). As a consequence, populations of Native Americans in emerging Portuguese and Spanish colonies were soon decimated, and colonizers had to look for labour force elsewhere. During the 16th century Portuguese and Spanish traders deported into American colonies more than 196 thousands of slaves, primarily from Senegambian region and West Central Africa, the coast south of mouth of river Congo. By the end of the 16th century, Brazil received more than 34 thousands of African slaves, mainly in Pernambuco (more than 18 thousands) and Bahia (more than 5 thousands). At the same time, Spanish colonies in Central America received about 50 thousands of slaves and other Spanish colonies in Americas another 119 thousands of enslaved Africans (Eltis & Richardson 2010).

In mid-17th century sugar production in Caribbean took off. British founded their plantations in Barbados (1624) and Jamaica (1655). From 1641 until the end of 17th century, British disembarked almost 400 thousands of African slaves in British Caribbean. At the end of the 17th century, a substantial amount of gold was discovered in Brazilian provinces of Minas Gerais and Goias. This catalyzed even faster the expansion of the slave trade. Slavers started to exploit more areas in African coast: Angola, Bight of Biafra and Bight of Benin (Eltis 2008, **Figure 12**).

1.5.3 Triangular trade and second Atlantic system

At the end of the 16 th century, a new model of trade emerged in the Atlantic. It was called triangular trade because of its typical three legs: Europe or New England as a source of manufactured goods, African coast as a source of slaves and American colonies as a source of exotic goods like molasse, coffee, cocoa and others. Given the state of technology in the 17th and 18th centuries all vessels were propelled solely by wind.

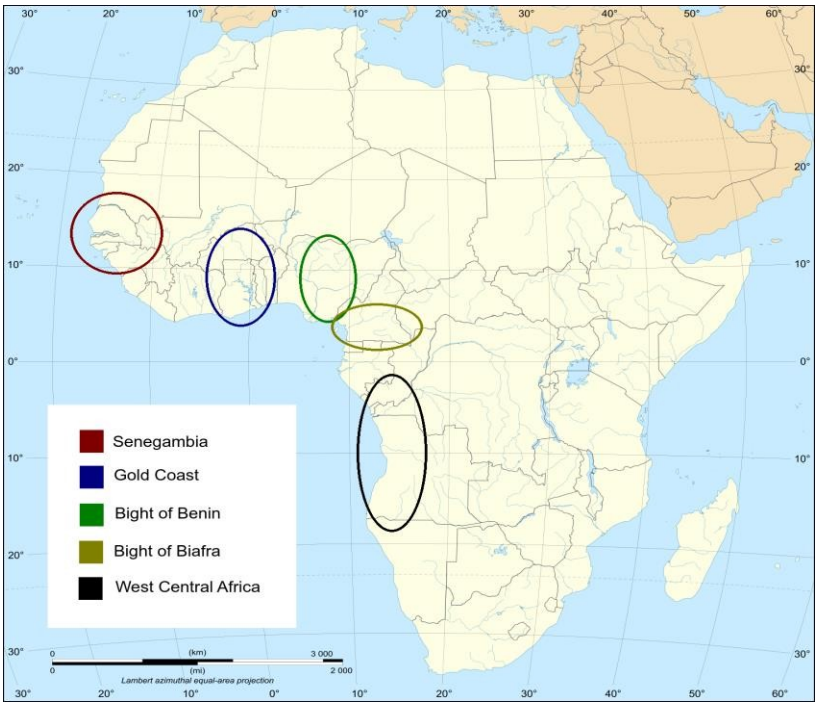

Figure 12 – Major embarkation regions in West Africa. The colored circles indicate broad embarkation regions frequented by slave traders. Based on information from Transatlantic Slave Trade Database (Eltis & Richardson 2010).

The navigation in the triangular trade system was only possible thanks to Trade Winds, a pattern of prevailing winds in the Atlantic: easterly winds in latitudes near equator, and westerly winds in latitudes north or south of 30°parallel. A typical voyage of a vessel in triangular trade started in a European port, carrying guns, ammunition, alcohol, beads and other goods to be traded for slaves. Upon arrival to the African coast, these goods were exchanged for slaves, which were then transported to American colonies. The voyage of slaves from Africa to Americas, often referred as the Middle passage, was extremely harsh. Due to the unpredictable nature of weather in the Atlantic, the duration of the middle passage was highly variable, typically from one to six months. As the ship owners tried to maximize the profit by packing the ships with as many enslaved people as possible, transported people had extremely small space to live on, which in many cases led to disastrous mortality rates during the crossing. The average mortality rate ranged between 9.5% (Spanish ships 1830-1867) and 29.8% (Spanish ships 1590-1699). However, it has been estimated that up to 70% of mortality could have occurred in Africa, during the kidnapping and transportation to the ports (Klein 1999).

Despite the fact that all European naval powers entered in TAST by the end of the 17th century, only during the 18th century the trade reached an enormous scale. In total, six naval powers entered the business along with the well-established Portuguese and Spanish: British, Dutch, French and Danish.

1.5.4 Great Britain in the Atlantic Slave Trade

British involvement in slave trade increased rapidly after introduction of sugar cane to Barbados by Dutch traders. In 1672 the Royal African Company was founded in London, and until 1698 London port had monopoly for slave trade in Britain. Later the operation centers of British slave trade moved to the ports of Liverpool and Bristol (Thomas 1997). During the 18th century, British ships carried over 2.5 million Africans across the Atlantic, most of them to Jamaica (close to one million) and to Barbados (around 328 thousands). Other islands in British Caribbean received between 100 and 150 thousand slaves: St. Kitts (149,600), Antigua (139,500), Grenada (137,100) and Dominica (114,100). British also shipped around 300 thousands of Africans to their North American colonies, primarily Carolina, Georgia and Virginia (Eltis 2008; Eltis & Richardson 2010). Trading to North America declined after the United States of America declared independence in 1776, but slave trade in British Caribbean continued until abolition in 1807. The principal regions of embarkation frequented by British were Gold Coast (ports of Elmina and Accra) with almost 645 thousands of embarked slaves, and Bight of Biafra (ports of Bonny and Calabar), accounting for more than 890 thousands. Another major region was

West Central Africa, where almost half-million of enslaved Africans embarked British vessels during the peak phase of the Atlantic Slave Trade in the 18th century.

1.5.5 France in the Atlantic Slave Trade

French involvement in the Atlantic Slave trade began shortly after France started the colonization of Caribbean. In 1635 Pierre Belain d'Esnambuc founded the first permanent colony on Martinique. Later, when British pushed French out of Saint Kitts and Nevis, French colonizers turned to Saint Lucia (1643) and Guadeloupe (1674). Until the end of the 17th century, French slave trade was relatively small, compared to Portuguese, British or even Dutch. During the last decade of the 17th century, when both British and Portuguese vessels embarked over 100 thousands Africans, the scale of French slave trade reached only about 10% of British and less than 7% of Portuguese. During the first three decades of the 18th century, French Atlantic Slave Trade ran up to 74 thousands of slaves between 1721-1730 and continued growing up to ~280 thousands in 1781-1790, when French slave trade saw its peak. The vast majority of slaves from French vessels were disembarked in French Caribbean, predominantly to Saint-Domingue and Martinique, which together accounted for more than 1.1 million out of 1.3 million of slaves transported to the New World during the whole French slave trade (Eltis & Richardson 2010).

Most of the French slavers expeditions (over 40%) started in the port of Nantes, a rich merchant city in Brittany, situated on river Loire approximately 50 km from the mouth in the Atlantic Ocean. Other important ports involved were Le Havre, La Rochelle, Bordeaux, St. Malo and Honfleur (Geggus 2001). French slave traders bought the slaves basically in all West African regions, but mainly in West Central Africa and Bight of Benin.

Similarly to other Caribbean colonies, enslaved Africans on French islands cultivated and processed sugar cane, coffee, cacao, indigo, tobacco and cotton. At the end of the 18th century, French slave trade started to decline, while the abolishment movement in France was growing stronger. Low profitability of slave trade and intense opposition against slavery in French society led to abolition of slavery in all of its possessions in 1794. However, this abolition did not last long, as Napoleon reintroduced slavery in sugar-growing colonies in 1802.

1.5.6 Netherlands in the Atlantic Slave Trade

Dutch involvement in the Atlantic Slave Trade started already at the end of the 16th century, when vessels under the Dutch flag shipped slaves to Brazil and Spanish colonies, however at

a smaller scale when compared to other powers. With the onset of Dutch colonies in Caribbean, the destination of Dutch slave vessels shifted towards Aruba, Curaçao and Sint Maarten. In the 18th century, the predominant markets of Dutch slave trading became Surinam and Guyana. Based on archive documents, Dutch slave trade affected more than 550 thousand Africans, who were embarked on Dutch ships. The vast majority of Dutch slave expeditions took place in the period between 1660 and 1790. Slave trade was abolished in the Kingdom of Netherlands only in 1863, being one of the last countries in Europe to do so. Furthermore, the adopted law of abolishment implied 10 years of transformation period, therefore only after 1873 all Dutch slaves were freed.

1.5.7 Denmark in the Atlantic Slave Trade

Given the small extension of Danish colonies in the New World, consisting of only three islands (Saint Thomas, Saint John and Saint Croix), the Danish involvement in the slave trade was marginal. Documented expeditions of Danish slave vessels account for 111 thousand enslaved Africans. Slavery in Danish colonies was abolished in 1848.

1.5.8 Abolition of the Slave Trade

Towards the end of the 18th century, profitability of the triangular trade declined gradually. There was an over-supply of sugar in the world, as well as of other goods, which were cheaper and easier to be produced in colonies outside the New World and without slave labour. In April 1791, a large slave rebellion began in the French colony of Saint-Domingue. The rebellion lasted over a decade, until in November 1803 rebels defeated French army and in the next year Saint-Domingue became the independent republic of Haiti. These events provoked heated dispute in colonial countries about the morality of slavery. In the center of this debate was an abolishment movement, which was significantly inspired by the ideas of humanity and universal human rights. Starting with Great Britain in 1807, all colonial powers proceeded to abolition of slave trade and slavery in the course of the 19th century. After the abolition in Great Britain, British navy pointed a fleet of ships to patrol in West Africa and enforce shut down of the slave trade. During the first half of the 19th century, Great Britain signed an abolition treaty with Portugal, Denmark, France and Netherlands. Finally, Brazil abolished slavery by the Golden Law, which came into effect on 13th May 1888, with immediate effect, although some illegal slave trade continued for several years after the abolition (Eltis & Walvin 1981).

1.6 Interdisciplinary study of the Trans-Atlantic Slave Trade

The first source of information available for the study of TAST was archive documents. All European slave traders kept more or less detailed records about the slave trade, which usually contain information about the number of bought slaves, the place of purchase, number of males, females and children, port of embarkation, flag under which the vessel set on sail across the Atlantic and where the slaves were disembarked and sold, length of the voyage and mortality during the voyage. These archive records provide relatively accurate information on the destination of the enslaved Africans, but contain only very little information on their origin, restricted to a broad geographic region, like Bight of Biafra or Golden Coast, or eventually to a particular port. As Africans were often captured far from the coast, their origin and ethnicity are in many cases unknown. In addition, mortality of the captured slaves during their transport to the coast was also never recoded, although some scholars estimate that it was even higher than the mortality during the harsh Middle Passage.

Because the amount of information that historical archives can provide is very limited, investigation of TAST requires interdisciplinary approaches to address questions about ethnic origin of slaves, their culture and habits and their health.

The first attempts to elucidate the ethnic origin of the first generation of enslaved Africans were made by using the isotope analysis (Cox & Sealy 1997; Price et al. 2006). Isotopes are variants of particular chemical elements that differ in number of neutrons. Elements like strontium, oxygen, carbon and nitrogen form stable isotopes, which in nature occur in different ratios often determined by geological conditions. Therefore, the ratio of stable isotopes can be used as a signature typical of a particular geographic region. As living organisms incorporate elements from the environment into their bodies, their biological materials contain the same isotope ratio as the environment they live in and can be recovered even after their. Isotope ratio recovered from biological material, e.g. skeletal remains of enslaved Africans, can be compared to the isotope map of a particular region in order to determine the possible origin of the investigated sample.

In the context of TAST, this strategy has been successfully applied to skeletal remains of enslaved Africans unearthed in Barbados. In three out of 25 samples, the isotope profiles obtained from skeletons yielded results which were not consistent with isotope ratios found in Caribbean, suggesting that these three skeletons represent the first generation of slaves born in Africa, albeit their exact origin could not be determined (Schroeder et al. 2009).

Improvements in capture techniques of ancient DNA and next-generation sequencing allowed

investigation of genetic background of enslaved Africans on genome-wide level. This can be done by statistical comparison of set of markers typed in ancient DNA samples to the reference data set of African populations. If the source population is present in the reference data set, the investigated sample should cluster with the parent population. The study of ancient DNA in the context of TAST is peculiar because of the generally low quality of DNA caused by post-mortem degradation. The rate of DNA degradation is largely determined by environmental factors: in the wet and hot climate of equatorial Africa and Caribbean, the preservation of DNA is much worse than in dry and cool conditions of higher latitudes. However, because the genetic landscape of African populations is highly stratified, even an unobtrusive number of successfully typed markers can eventually provide a good spatial resolution. This approach was successfully used to determine the origin of three enslaved Africans from the 17th century unearthed in the Caribbean island of Sint Maarten. By using the reference data set compiled of published West-African samples, two out of three samples were assigned to a particular source population (Schroeder et al. 2015).

1.7 Genetic legacy of Trans-Atlantic Slave Trade in clinical context

1.7.1 Health related factors of African ancestry

Genetic ancestry background plays an important role in health of every individual, as expressed in susceptibility and resistance to diseases. Health implications arising from African genetic ancestry are extremely relevant for the public health sector of countries with populations of African descent: African Americans, African Caribbeans and African Brazilians.

Although a considerable amount of the disparity observed in health status between African Americans and European Americans can be attributed to socioeconomic factors (Schulz et al. 2000), there are several diseases and health conditions affecting African Americans significantly more frequent than in other ethnic groups in the population, even after correction for socioeconomic factors. Complex diseases prevalent in African Americans, like hypertension, type 2 diabetes, progressive kidney failure and blood disorders, have known genetic component, and risk alleles of these diseases are frequently linked to African ancestry (Genovese et al. 2010; Parsa et al. 2013). On the contrary, African ancestry seems to provide protection against some infectious diseases (Chacón-Duque et al. 2014).

1.7.1.1 Kidney diseases in African Americans

Chronic Kidney Disease (CKD) is a progressive loss of kidney function characterized by reduced glomerular filtration and/or increased urinary albumin excretion (Jha et al. 2013). In the United States, this disease affects African Americans approximately two times more than European Americans (Cowie et al. 1989, Tell et al. 1996). Genome-wide scan for association tagged two risk alleles in the coding region of Apolipoprotein 1 gene (*APOL1*): rs73885319, a SNP of A for G, designed as G1 (linked with rs60910145); and rs71785313, designed as G2, a 6-bp deletion causing removal of two amino acids (Genovese et al. 2010). Both of these two derived variants G1 and G2 are present only in Sub-Saharan populations, where G2 variant presents a 3-8% frequency in West Central Africa (Ko et al. 2013; **Figure 13**).

Further clinical studies proved that G1 and G2 variants not only affect the susceptibility to CKD, but also influence its progression. Patients with two risk alleles (high risk group) progressed towards the kidney failure (primary renal outcome) faster than patients with only one or no risk allele (low risk group). After 10 years of monitoring, roughly 70% of high risk patients came to renal outcome, compared with40% of patients in the low risk group (Parsa et al. 2013).

A high frequency of deleterious variants in a relatively confined geographic area strongly

suggests that variants are maintained in the population by balancing selection, thus these variants could confer an evolutionary advantage, which is strong enough to counterbalance increased risk of CKD. This hypothesis has been corroborated by analysis of haplotype structure in *APOL1* gene: both G1 and G2 haplotypes are significantly longer than wild type haplotypes and present high scores in selection tests (Karlsson et al. 2014).

The competitive advantage of the derived forms of apoliprotein was demonstrated by *in vitro* experiments, which proved the ability of the derived Apol-1 protein to lyse the protozoan parasite *Trypanosma brucei rhodesiense*, a causative agent of sleeping sickness. Apolipoprotein 1 is the trypanolitic factor of human serum, responsible for protection against trypanosome parasites. The molecular basis of this protection lies in its ability to form pores in the lysosomal membrane of the parasite and induce influx of chloride, followed by osmotic swelling and consequent lysis of the trypanosome (Pérez-Morga et al. 2005). However, the trypanosome subspecies *T. b. rhodesiense* and *T. b. gambiense* adapted to humans and express the serum resistance associated (SRA) protein that binds to terminal helix of wild type Apol-1 protein and inhibits its protective activity (Vanhamme et al. 2003). Nevertheless, binding of SRA to Apol-1 can be disrupted, if terminal helix of Apol-1 is altered by deletions or mutations, which is the case of G1 and G2 versions of the protein. These mutated proteins retain the lytic capacity even in human adapted forms *T. b. rhodesiense* and *T. b. gambiense* (Genovese et al. 2010).

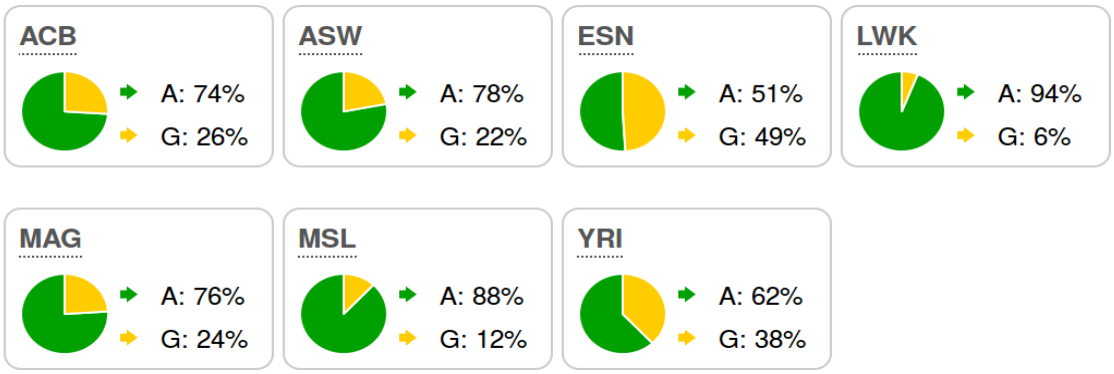

Figure 13 - Frequency of derived G1 allele in African populations. ACB: Barbados, ASW: African American, ESN: Esan from Nigeria, LWK: Luhya from Kenya, MAG: Mandinka from Senegal, MSL: Mende from Sierra Leone, YRI: Yoruba from Nigeria. Pie charts downloaded from www.ensembl.org (Cunningham et al. 2014).

Figure 7 - Distribution of global Dengue risk. Adapted from Global Strategy for Dengue Prevention and Control (WHO 2012).**Figure 8 - Frequency of derived G1 allele in African populations.** ACB: Barbados, ASW: African American, ESN: Esan from Nigeria, LWK: Luhya from Kenya, MAG: Mandinka from Senegal, MSL: Mende from Sierra Leone, YRI: Yoruba from Nigeria. Pie charts downloaded from www.ensembl.org (Cunningham et al. 2014).

1.7.1.2 Type 2 diabetes

Diabetes mellitus type 2 is a metabolic disorder manifested in elevated blood glucose level (hyperglycemia), caused either by relative shortage of insulin supply due to dysfunction of β-cells or by resistance of cells to insulin. Untreated diabetes can lead to impaired vision, kidney and cardiovascular disorders and sexual dysfunctions (Lin & Sun 2010). Although obesity is the main risk factor of type 2 diabetes, trials on homozygotic twins suggest a strong heritability: coincidence of type 2 diabetes among identical twins exceeds 90% (Melmed & Conn 2005).

Type 2 diabetes in African Americans is approximately doubled compared to European Americans (Maskarinec et al. 2009). Despite the extensive effort, only a few genetic variants significantly associated with type 2 diabetes were identified in African American: one intergenic SNP (rs7560163) between *RND3* and *RBM43* genes (Palmer et al. 2012); one SNP (rs7903146) in *TCF7L2* gene (Cooke et al. 2012, Long et al. 2012) and several SNPs in *HLA-B* and *IGF2* genes (Ng et al. 2014).

1.7.1.3 Hypertension

Arterial hypertension is a persistent medical condition characterized by increased blood pressure in arteries. In adults, this means arterial blood pressure higher than 140 mmHg for systolic and 90 mmHg for diastolyc. Hypertension is an important risk factor in a number of disorders, in particular hypertensive heart disease, CKD, stroke, etc.

Prevalence of hypertension in African Americans, age adjusted, is approximately 39%, compared to 27.8% in Mexican Americans or 28.5% in non-Hispanic white Americans (Ong et al. 2007). Several GWAS aimed to disentangle between genetic variants responsible for the higher incidence of hypertension in African Americans, however few results were successfully replicated in an independent sample. Admixture mapping in 6303 unrelated African Americans followed by GWAS replication identified a variant in *NPR3* gene associated with hypertension (Zhu et al. 2011). In addition, a large meta-analysis of previous GWAS studies revealed other five loci: *EVX1-HOXA*, *ULK4*, *RSPO3*, *PLEKHG1*, and *SOX6*, all of them involved in the nitric oxide signaling pathway (Franceschini et al. 2013). This pathway plays a role in various processes linked to hypertension, such as heart contraction, vasodilatation and endothelian function (Saraiva & Hare 2006).

1.7.1.4 Blood disorders

Higher prevalence of blood disorder among people of African ancestry can be explained by the protective effect against malaria which these disorders often provide (Kwiatkowski 2005).

Thalassemia and sickle cell anaemia are inherited genetic blood disorders which affect haemoglobin formation and, in this way, prevent malaria parasite from successful completion of its reproductive cycle inside the erythrocyte. The causative genetic variants are maintained by balancing selection in regions endemic for malaria parasites. However, due to extensive migrations related to TAST thalassemia and sickle cell anaemia are present in all populations with West African ancestry (Steinberg et al. 2009).

1.7.1.4.1 β-thalassemia

β-thalassemia is a blood disorder defined as an insufficient synthesis of haemoglobin β-chains typically leading to anaemia. Thalassemia is determined by mutation in *HBB* gene and is inherited in Mendelian autosomal recessive fashion (Cao & Galanello 2010). The heterozygous state with one functioning *HBB* gene is referred as β + -thalassemia (β-thalassemia minor). The heterozygous phenotype of varies from asymptomatic carriers to severe anemic individuals (Sheiner et al. 2004). Homozygous individuals for β-thalassemia trait (β-thalassemia major) often suffer from severe anaemia and depend on regular blood transfusion. Additionally, frequent blood transfusions might cause problems with iron overload and consequent enlargement of the spleen (Wintrobe & Greer 2009).

1.7.1.4.2 Sickle cell anaemia

Sickle cell anaemia is caused by the substitution of glutamic acid by valine in haemoglobin β-chain. Mutated β globine causes sickle-shaped erythrocytes with impaired capability of flowing through veins. Sickle cell trait (HbS) in heterozygous state usually causes only mild anaemia while providing partial resistance against malaria parasite *P. falciparum*. On the contrary, individuals homozygous for HbS trait suffer from severe anaemia that considerable shortens life expectancy (Rees et al. 2010). Prevalence of sickle cell anaemia in African Americans is approximately 0.02% and around 8% of African Americans bear HbS trait, hence are heterozygous carriers (Ojodu et al. 2014).

1.7.2 Case study: Dengue fever

Dengue is a tropical disease spread by mosquitoes of the genus *Aedes.* Symptoms of Dengue fever (DF) are headache, fever and pain in muscles, joints and bones. This is often accompanied by typical skin rash. Most of the patients recover from Dengue infection without serious health problems. Around 80% of infected people are asymptomatic or exhibit only very mild symptoms (WHO 2009; Reiter 2010). In some cases the Dengue infection progresses into dengue hemorrhagic fever (DHF), characterized by blood plasma leakage and low levels of

blood platelets. Another life-threatening condition induced by Dengue infection is a Dengue shock syndrome (DSS) associated with low blood pressure. If adequate treatment is provided, the fatality rate is less than 1% (WHO 2009), without the treatment the fatality is estimated to 1%-5% (Ranjit & Kissoon 2011).

1.7.2.1 Virus

The causative agent of Dengue fever is the Dengue virus (DENV). DENV is a single-stranded RNA virus belonging to the *Flaviviridae* familiy. Based on the envelope protein, DENV is subclassified into four DENV serotypes (DENV 1-4) and a fifth serotype has been discovered in 2013 (Normile 2013). Infection with one serotype provides life-long resistance to that particular serotype, but not to other serotypes. In fact, secondary infection with a different serotype is a risk factor for developing the life-threatening forms of Dengue fever (Rodenhuis-Zybert et al. 2010).

1.7.2.2 Vector

Dengue is transmitted by mosquitoes of the genus *Aedes*, most often by species *A. aegypti* and *A. albopictus* (Gubler 2004). Mosquitoes of the genus *Aedes* usually live in wet and warm tropical habitats in latitudes between 35N and 35S parallel, and below 1000 m altitude. *Aedes* mosquitoes are day-biting species, and besides Dengue fever, serve as vector for yellow fever, chikungunya and West Nile virus. Mosquitoes from the genus *Aedes* are characterized by white markings on body and legs.

1.7.2.3 Epidemiology

The Dengue fever is endemic in 110 countries in Central and South America, Africa, South and Southeast Asia (Ranjit & Kissoon 2011). Within the last decades, the incidence of Dengue increased dramatically. While there were only several thousands of Dengue cases reported in 1960s, in 2010 WHO received reports about more than 2.2 million of Dengue fever cases (WHO 2012). This increase can be attributed to the spread of Dengue vector related to climate change, globalization and urbanization, as the Dengue infections are most common in urban environment (Gubler 2004).

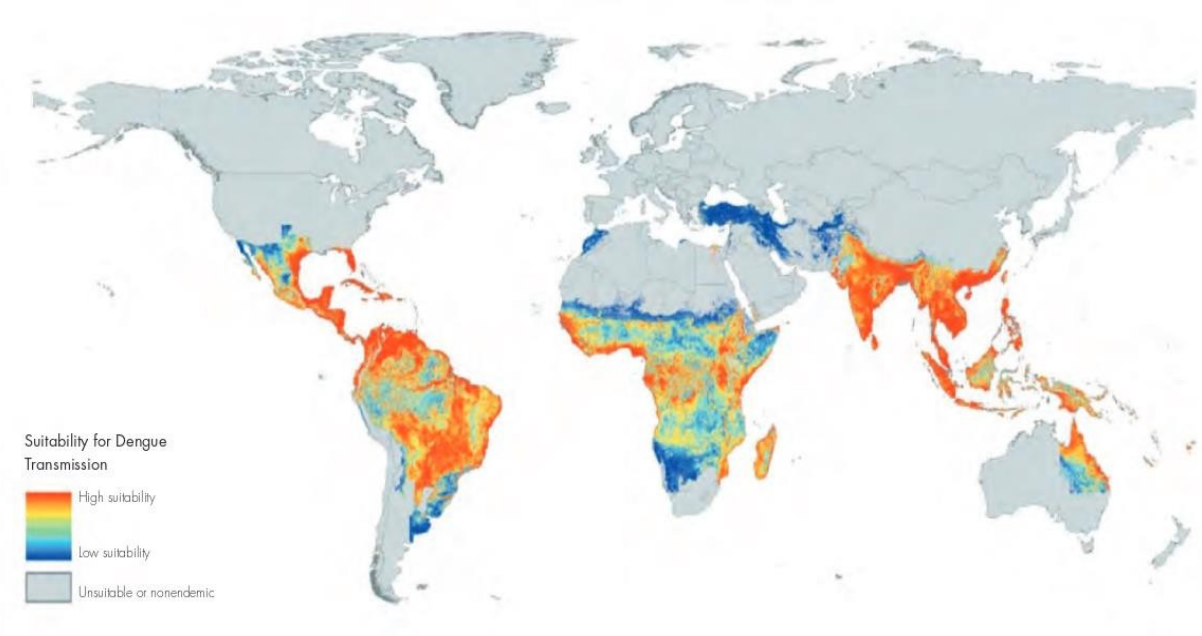

Figure 14 - Distribution of global Dengue risk. Adapted from Global Strategy for Dengue Prevention and Control (WHO 2012).

1.7.2.4 Role of African ancestry in Dengue

Medical reports of the Dengue outbreak at Cuba in 1981 pointed out that ethnicity might play a role in the outcome of the disease, because people with darker skin showed better resistance to the disease. However, until now this hypothesis was tested only in two genetic studies. The first attempt was conducted in 236 DF and 50 DHF patients from Brazil, through single locus analysis of 593 SNPs targeting genes in the IFN response pathway, which is involved in arboviral resistance in mice. This study found association between the susceptible phenotype and non-African ancestry on *JAK1* gene, possibly, interacting with IFN response pathway (Silva et al. 2010).

Another study targeted 30 ancestry informative markers in Colombian dengue patients and found statistically significant protective effect of African ancestry, however due to low resolution it could not provide answers on the mechanism of protection (Chacón-Duque et al. 2014).

1.8 History of Arab slave trade

Arab Slave Trade in East Africa was a part of a much larger slave trading system in the Muslim world. It started in the 7th century with Muslim conquests. At its largest extent in the 8th century during the Umayyad caliphate, the Muslim Empire spread throughout North Africa, Iberian Peninsula, Arabian Peninsula, Near East and parts of Central Asia. Because the Muslim law sharia forbids enslavement of fellow Muslims, Arab slave traders raided areas outside the Muslim Empire, primarily European coasts (slaves called Saqualiba) and Sub-Saharan Africa (slaves called Zanj). Because the Arab Slave Trade lasted more than one millennium and the number of written records about the Arab Slave Trade is very limited, estimates on the scope of the trade vary between authors. Including the Trans-Saharan slave trade, estimates start at 8 million Africans and go as high as 25 million Africans enslaved between 7th and 19th centuries by Arab slavers (Gordon 1989; Smith 1999).

Arab slave trade in Africa had two main branches: the Trans-Saharan and the Indian Ocean. The Trans-Saharan slave trade operated in West Africa, in particular in the Kingdom of Mali. Slaves were gathered in cities like Tibmuktu, Gao and Djemne. From there the slaves were transported across the Sahara to the markets in Maghreb: Marrakech, Algiers, Tunis and others (**Figure 15**).

In East Africa, Arab slave traders captured people along the East African coast and shipped them into countries on the coast of the Indian Ocean. Slaves were used for a considerable number of tasks. A large number of slaves was used for agriculture and army, but the most important trade was of women. This is apparent in the maternal genetic pool of Arab populations: African-specific mtDNA haplogroups are present in high frequencies in Arab populations, especially in Yemen, where it reaches 34%; on the other hand, mtDNA L haplogroups are not frequent in non-Arabic Near Eastern populations (Richards et al. 2003).

Dating of the Arab Slave Trade from genetic data is particularly complicated because of the geographic proximity in Horn of Africa and due to long-standing admixture between populations in this region. As we discuss later, LD based methods like ROLLOFF (Moorjani et al. 2011) tend to detect only the most recent admixture and omit the more ancient events. New recently developed method by Hellenthal et al. (2014) based on chromosome painting and haplotype analysis can better estimate admixture time and even distinguish several migration waves, however in case of African admixture in Arabian peninsula detects only the Arabian Slave Trade in period between 890 and 1750 CE, which is in line with the historical evidence (Manning 1990).

The genetic legacy of the Arab Slave Trade has also implications for the health of present day Arabian populations. The sex-biased gene flow from sub-Saharan Africa into Arabian Peninsula is a reason for high frequency of blood disorders in the Arabian Peninsula, in particular X chromosome linked G6PD deficiency, the cause of fava anaemia. Blood disorders like G6PD deficiency or various haemoglobin malformations provide a certain degree of protection against malaria and are very frequent in Sub-Saharan Africa (Kwiatkowski 2005; Karlsson et al. 2014).

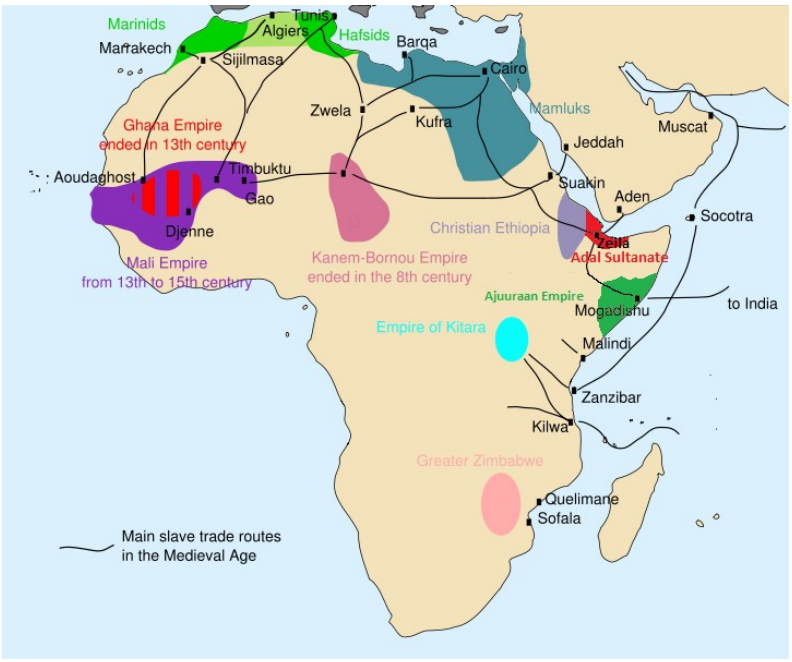

Figure 15 – Routes of Arab Slave Trade. Adapted from “African slave trade" by Runehelmet derived from Aliesin - File:Traite_musulmane_medievale.svg. Licensed under CC BY-SA 3.0 via Commons license.

2 Aims

A good understanding of population structure and evolutionary forces that acted upon ancestral and descendant populations is needed to investigate the genetic legacies of Africa's largest slave trading systems (TAST and Arab Slave Trade). The work presented in this thesis aims to (i) describe the genetic structure and positive selection across the Sahel belt, whose western populations contributed massively to TAST while the eastern populations were the ancestors of the Arab Slave Trade; (ii) investigate how the African genetic background can contribute to protection against complex diseases in admixed descendant populations, by using as a model Dengue hemorrhagic fever in the Cuban population; and (iii) provide genetic evidence and date estimation of demographic events in complex African and non-African admixture scenarios in Arabia.

Specific aims of this work:

1. **Characterize patterns of admixture and positive selection of populations in Africa's most important migration corridor.** Sahel belt has a complex history of human migrations, admixture and evolutionary adaptation. The population structure of Sahel was inferred from 2.5 million genome-wide SNPs characterized in 161 individuals from 13 populations by using ADMIXTURE and PCA. Signals of positive selection were also investigated by using haplotype-based selection measures (iHS and XPEHH) and contextualized in terms of metabolic pathways to elucidate possible biological adaptations.

2. **Investigate the role of African ancestry in resistance and susceptibility to complex diseases in admixed descendant populations of the slave trade, by using the model of Dengue hemorrhagic fever in admixed Cubans.** Although the protective effect of African ancestry against the Dengue hemorrhagic fever has been proposed by health professionals in the Caribbean region based on empirical experience, the genetic basis of this phenomenon was not known. ADMIXTURE analysis was used to evaluate the hypothesis of genetic African protection against the hemorrhagic phenotype in Cuba, while fine-matched genome-wide association study and ancestry mapping were applied in order to identify the candidate genes conferring the African protection. Gene expression of the identified genes was evaluated in Cuban patients and in available transcriptome data from a Thai cohort.

3. **Shed light on the Arab Slave Trade impact in the complex admixture scenario in Arabia.** Since Pleistocene, the population genetics of Arabian Peninsula has been shaped by episodes of gene flow between Africa and Eurasia resulting from climatic oscillations, Neolithic expansions and slave trade. We compared the power between mitochondrial and autosomal based analysis in disentangling between and dating diverse migration events.

3 Papers

Paper I - Extensive admixture and selective pressure across the Sahel Belt.

Paper II - *OSBPL10, RXRA* and lipid metabolism confer African-ancestry protection against haemorrhagic fever in admixed Cubans.

Paper III - Genetic Stratigraphy of Key Demographic Events in Arabia.

3.1 Paper I

Extensive admixture and selective pressure across the Sahel Belt.

Genome biology and evolution, *7*(12), 3484-3495.

GBE

Extensive Admixture and Selective Pressure Across the Sahel Belt

Petr Triska[1,2,3], Pedro Soares[2,4], Etienne Patin[5,6],Veronica Fernandes[1,2], Viktor Cerny[7], and Luisa Pereira[1,2,8,*]

[1]Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Porto, Portugal

[2]Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto, Portugal

[3]Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto (ICBAS), Porto, Portugal

[4]Department of Biology, CBMA (Centre of Molecular and Environmental Biology), University of Minho, Braga, Portugal

[5]Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, France

[6]Centre National de la Recherche Scientifique, Paris, France

[7]Archaeogenetics Laboratory, Institute of Archaeology of the Academy of Sciences of the Czech Republic, Prague, Czech Republic

[8]Faculdade de Medicina da Universidade do Porto, Porto, Portugal

*Corresponding author: E-mail: lpereira@ipatimup.pt.

Date deposition: This project has been deposited at the European Genome-Phenome Archive under the accession EGAS00001001610.

Accepted: November 21, 2015

Abstract

Genome-wide studies of African populations have the potential to reveal powerful insights into the evolution of our species, as these diverse populations have been exposed to intense selective pressures imposed by infectious diseases, diet, and environmental factors. Within Africa, the Sahel Belt extensively overlaps the geographical center of several endemic infections such as malaria, trypanosomiasis, meningitis, and hemorrhagic fevers. We screened 2.5 million single nucleotide polymorphisms in 161 individuals from 13 Sahelian populations, which together with published data cover Western, Central, and Eastern Sahel, and include both nomadic and sedentary groups. We confirmed the role of this Belt as a main corridor for human migrations across the continent. Strong admixture was observed in both Central and Eastern Sahelian populations, with North Africans and Near Eastern/Arabians, respectively, but it was inexistent in Western Sahelian populations. Genome-wide local ancestry inference in admixed Sahelian populations revealed several candidate regions that were significantly enriched for non-autochthonous haplotypes, and many showed to be under positive selection. The *DARC* gene region in Arabs and Nubians was enriched for African ancestry, whereas the *RAB3GAP1/LCT/MCM6* region in Oromo, the *TAS2R* gene family in Fulani, and the *ALMS1/NAT8* in Turkana and Samburu were enriched for non-African ancestry. Signals of positive selection varied in terms of geographic amplitude. Some genomic regions were selected across the Belt, the most striking example being the malaria-related *DARC* gene. Others were Western-specific (oxytocin, calcium, and heart pathways), Eastern-specific (lipid pathways), or even population-restricted (*TAS2R* genes in Fulani, which may reflect sexual selection).

Key words: genome-wide diversity, admixture, selection, Sahel.

Introduction

Africa was the cradle of modern humans and the only inhabited continent for around two-thirds of their history, hence extant African populations harbor the greatest worldwide genetic diversity (Prugnolle et al. 2005; Soares et al. 2012; Rito et al. 2013). This rich genetic pool has been modeled by complex demographic (changes in population size, migration events, and admixture) and genetic (natural selection, recombination, and mutation) events. Genomic studies in African populations are therefore of paramount potential in revealing main aspects of human population history and genetic susceptibility to diseases. Before the advent of genome-wide studies (GW), a large screening in Africa consisting in 1,327 microsatellites showed that the population structure follows quite reasonably the self-described ethnic and linguistic groups, implying overall a large and subdivided population structure in Africa that was opposed in several populations by extensive mixture of ancestries owing to large-scale migration events that occurred throughout history (Tishkoff et al. 2009). First GWs focused on African-Americans and in their Western

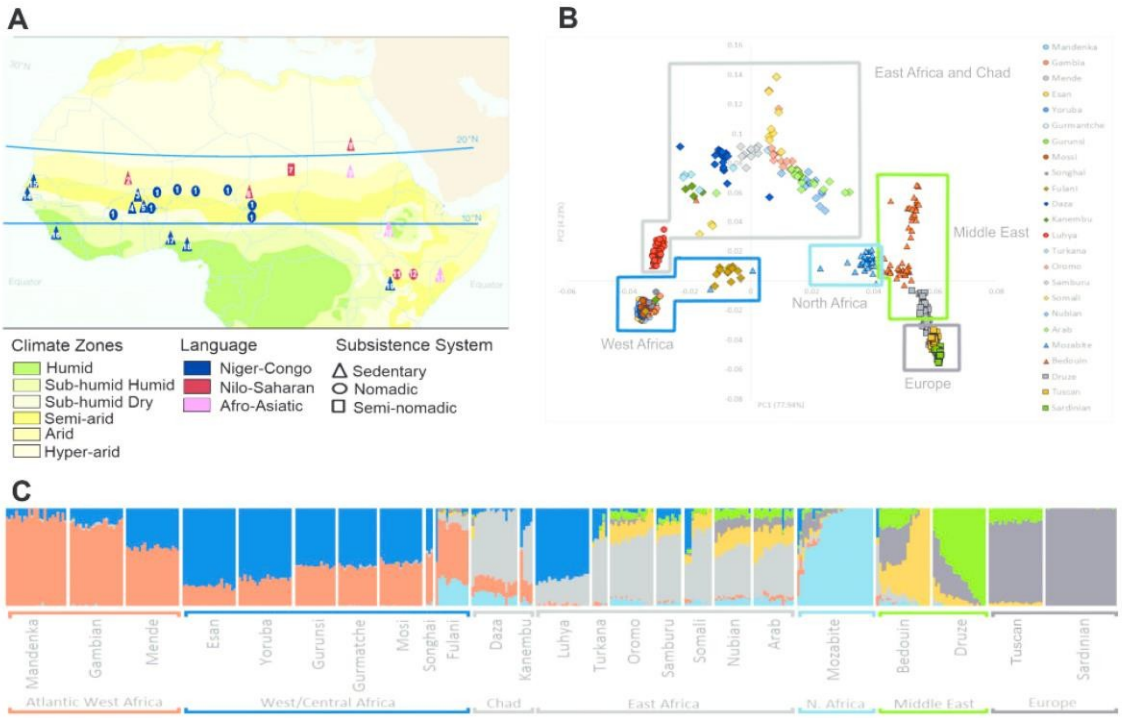

FIG. 1.—Location of studied samples and population structure across Sahel. (*A*) Geographic locations of the populations studied here, with subsistence system and family language affiliation identified. The colored zones indicate the current climate zones. Numbers 1 to 13 refer to groups studied here: 1 – Fulani; 2 – Songhai; 3 – Mossi; 4 – Gurunsi; 5 – Gurmantche; 6 – Kanembu; 7 – Daza; 8 – Nubians; 9 – Sudanese Arabs; 10 – Oromo; 11 – Samburu; 12 – Turkana; and 13 – Somali. Numbers 14 to 19 refer to groups from 1000 Genomes project and Li et al. (2008): 14 – Gambian in Western Division; 15 – Mandenka in Senegal; 16 – Mende in Sierra Leone; 17 – Yoruba in Ibadan, Nigeria; 18 – Esan in Nigeria; and 19 – Luhya from Kenya. (*B*) PCA1 versus PCA2. (*C*) Admixture analysis for K = 7 ancestral populations (each represented by a color). Each vertical line is an individual.

African ancestors (Smith et al. 2004; Zakharia et al. 2009), and even the 1,000 Genomes project (Abecasis et al. 2012) includes mainly Western African samples (four Western and one Eastern of Bantu/Western ancestry). Other GWs began to describe either population structure of hunter-gatherers as the Pygmies (Patin et al. 2014), Khoisan (Pickrell et al. 2012) and Khoisan-speaking Hadza and Sandawe from Tanzania (Pickrell et al. 2012), or admixture with Eurasians in East and South Africa (Pickrell et al. 2014). But it remains difficult to pursue GW in other African populations across the continent and to evaluate the genomic effects of local demographic and selective pressures taking into account the myriad of environments, climates, diets, lifestyles, and exposure to infectious diseases (Campbell and Tishkoff 2008; Teo et al. 2010).

Among the African regions, the Sahel is of particular importance because of its major role as a migration corridor (Newman 1995; Cerny et al. 2009; Pereira et al. 2010; Cerny et al. 2011; Soares et al. 2012). The Sahel Belt lies between the Sahara desert in the north and the tropical forests in the south (more or less between parallels 20°N and 10°N), lacks high mountains or other barriers, and constitutes a particular and very specific ecosystem (fig. 1*A*) made of semi-arid grasslands, savannas, steppes, and thorn shrublands. The eastern part of the Belt goes even below the equator, including Ethiopia, South Sudan, Somalia, and Kenya. Despite its overall aridity (UNEP 2008), the Sahel has important water resources, consisting in long river courses (Nile, Niger, and Senegal) and Lake Chad. Climatic changes have transformed Sahel's aridity along time, from a notoriously arid-uninhabitable desert during the Last Glacial Maximum to a humid-fertile landscape of lakes and savannah from 10 to 6 thousand years ago (ka) in the Holocene Climatic Optimum (Drake et al. 2011). These climatic oscillations and annual cycles imply population expansions and migrations when conditions are favorable, and bottlenecks and isolation in refugia when periods are more difficult.

Two sympatric lifestyles co-exist in the Sahel: nomadic pastoralists who find pasture during the short rainy season in southern Sahara, and sedentary farmers who settled in the more humid areas. Pastoralists raise livestock, which was probably introduced 8-5 ka in Africa from the Near East, following

the Nile Valley, and eventually they used the Sahel Belt as a corridor to reach Western Africa (Hanotte et al. 2002), leaving genetic evidences in the human mitochondrial DNA pool (Cerny et al. 2009). Unlike nomadic pastoralism, sedentary farming is a rather recent phenomenon in the area (Marshall and Hildebrand 2002). The largest pastoral nomadic group in the world is the Fulani ethnic group, living today in 17 African countries and amounting to almost 40 million people, of which one-third still preserves the nomadic way of life (Cerny et al. 2006). Their origin is uncertain, with one of the hypothesis stating that in ancient times, the Fulani people moved from Egypt/Ethiopia to Senegal and at the beginning of the 13th century, migrated southwards, where grazing was available (Steverding 2008). Also, northwards of Lake Chad, the seminomadic Daza people live in small hamlets concentrated in oases around the Lakes of Ounianga (Podgorna et al. 2013), where they cultivate dates and grain, and practice transhumance of camels and donkeys. Turkana and Samburu, who originated in Sudan (Spencer 1965; Lamphear 1988) and live currently in North Kenya, are dedicated to cattle pastoralism and a traditional way of life, similar to their neighbor Maasai. Oromo and Somali were also nomadic in the past, but currently they are mostly sedentary (Lewis 1999; Etefa 2012).

Domesticated animals, which are essential to the former and extant economic Sahelian subsistence, also play a significant role as reservoirs of human parasites (Wolfe et al. 2007). Combined with climate and environmental conditions, the close contact between humans and cattle has been contributing to the massive burden of endemic and epidemic diseases in the Sahel. The Sahelian countries have one of the highest under-five years' mortality rates, with the majority of deaths mainly caused by pneumonia, diarrhea, and malaria. The region experiences recurrent outbreaks of cholera, measles, meningitis, polio, and typhoid (WHO 2012). It is among the worldwide regions with highest burden of malaria, and it has been shown that the vector *Anopheles gambiae* can survive the long Sahelian dry-season by entering into diapause (Huestis and Lehmann 2014). The Sahel region is known as the meningitis belt, with person-to-person spread epidemics caused by the bacteria *Neisseria meningitides* (Molesworth et al. 2002). Another highly conditioning disease in the region is the human African trypanosomiasis or sleeping sickness (Steverding 2008), caused by *Trypanosoma brucei gambiense* (in Western and Central Africa) and *T. brucei rhodesiense* (in Eastern and Southern Africa). These protozoan parasites are transmitted to mammalian hosts by the bite of infected tsetse flies (*Glossina* sp.), whose range in Africa overlaps largely the Sahel Belt. Other neglected tropical diseases, including a large majority caused by helminth infections (such as hookworm, schistosomiasis, and lymphatic filariasis), are also endemic in this region (Hotez and Kamath 2009). Thus, the Sahel Belt is a zone of strong selective pressure acting upon the human population.

To better characterize the genetic structure and adaptive history of Sahelian populations, we genotyped 2.5 million single nucleotide polymorphisms (SNPs; Human Omni 2.5 DNA Bead Chip, Illumina) in 13 populations ($n = 161$ individuals – fig. 1*A* and supplementary table S1, Supplementary Material 1 online) from the interior western region (Burkina), the Chad Basin, and across the Eastern side of the Sahelian zone. These populations are from diverse linguistic and subsistence system affiliations, and have extensively escaped the influence of the Bantu migration, which contributed greatly to homogenize the population diversity in the bulk of sub-Saharan Africa below the Belt (Gurdasani et al. 2015; Silva et al. 2015). We first evaluated the population structure in the Belt and then inferred candidate selected genomic regions by using two complementary haplotype-based tests, integrated haplotype score (iHS; Voight et al. 2006) and cross population extended haplotype homozygosity (XP-EHH; Sabeti et al. 2007). iHS has good power to detect selective sweeps (consisting in the fixation of a beneficial mutation) at moderate frequency (50–80%), and XP-EHH is most powerful for selective sweeps above 80% frequency (Voight et al. 2006; Sabeti et al. 2007). Notwithstanding the difficulty in proving that these candidate regions in fact reflect the action of positive selection (Hernandez et al. 2011), they display an outlier pattern of diversity that is consistent with positive selection and are enriched in true positives. The level of resolution of our study allowed us to get informative insights into the palimpsest of genome-wide candidate selected regions across the Sahel, one of the worldwide zones most exposed to infection.

Material and Methods

Population Samples, Genome-wide Genotyping, and Published Data

DNA samples analyzed in this study were collected from 13 populations in eight African countries. Further information relative to these populations is provided in supplementary table S1 (Supplementary Material 1 online) and geographic location can be found in fig. 1*A*. This study obtained ethical approval from the Ethics Committee of the University of Porto, Portugal (17/CEUP/2012). A total of 171 individuals were genotyped for the Illumina Human Omni 2.5 DNA Bead Chip, containing approximately 2.5 million SNPs. Nine samples from the Turkana population were excluded from the analysis, as they were closely related, and another Turkana was a clear ancestry outlier of the population (genetically close to Europeans). We ended up with 161 individuals. Quality control is summed up in supplementary fig. S1 (Supplementary Material 1 online), and a total of 2,247,183 SNPs passed quality control checking. To increase spatial resolution, we included 50 randomly selected unrelated individuals from relevant populations from the 1,000 Genomes Project (Abecasis et al. 2012), and from Li et al. (Li et al. 2008),

reported in supplementary table S2 (Supplementary Material 1 online). The extended low-density data set contained 370,470 SNPs and 732 individuals. The build used in all analyses was GRCh37.

Population Structure and Differentiation

Several population genetic analyses assume independent markers, so SNPs were pruned for pairwise linkage disequilibrium (LD) in PLINK (Purcell et al. 2007), by removing any SNP that had a $r^2 > 0.4$ with another SNP, within a 50-SNPs sliding window with step of 20 SNPs. Principal component analysis (PCA), which infers worldwide axes of human genetic variation from the allele frequencies of various populations, was carried out by using the *smartpca* tool, included in the EIGENSOFT package (Patterson et al. 2006). ADMIXTURE, which provides a maximum likelihood estimation of the population structure (Alexander et al. 2009), was run for several values (from 2 to 7) of clusters or ancestral populations, *K*. The optimal *K* was estimated through cross-validation of the logistic regression. Wright's F_{ST} metric was calculated using Vcf tools (Danecek et al. 2011). Details are provided in supplementary fig. S1 (Supplementary Material 1 online).

Local Ancestry Inference

We applied the RFMix algorithm (Maples et al. 2013), which uses a LD model between markers to infer ancestry for each segment of the genome between a mixture of two putative ancestral panels of haplotypes. We used as ancestral data sets the phased data from the 1,000 Genomes Database, the Italian sample representing southern European ancestry and Gambian or Luhya representing the African ancestry (the first when testing for Fulani and the later when analyzing Central and Eastern Sahelian populations). Italians are a good proxy population for the shared Mediterranean ancestry across southern Europe, North Africa, and the Near East/Arabian Peninsula, which is mixed with the sub-Saharan ancestry across the Sahel (Botigue et al. 2013). For comparison purposes, we assayed the effect of using Great Britain (GB) or Northern Europeans from Utah (CEU) as the non-African parental population in the Oromo data set, although by being derived North European populations, these are not geographically, historically, and anthropologically supported as good proxies for non-African admixture in the Sahel (Table S10). Some differences were observed; for instance, the main block on chromosome 2, in the region of the *LCT* gene, which has been mainly selected in North Europe, was identified in all three analyses, but the size was different: it was larger when using Italians (133,640,409-137,586,694, totaling 3,946,285 bp) than when using GB (133,346,735-135,210,390 – 1,863,655 bp) or CEU (133,346,735-134,727,858 – 1,381,123 bp). So, an input from a non-European population is confirmed in this region, although we do not know if the driver was *LCT* or another neighbor gene.

Our samples were phased in SHAPEIT v.2 (Delaneau et al. 2012) using HapMap reference panel and fine-scale genetic map. Information on ancestry was obtained for each locus along chromosomes for every individual, and these values were averaged in each population. The null hypothesis equates that the proportion of a certain parental ancestry in an admixed population will be equal across the genome. The test hypothesis is that a certain region of the genome may have a significantly higher proportion of a parental ancestry in comparison with its genome mean value, indicating the action of some event (as positive selection). To identify regions that have a significantly higher proportion of a given parental ancestry, we followed published studies (Bryc et al. 2010) considering only regions outside the range defined by the genome mean value for a certain ancestry ± 3SD. Genes identified as statistically significantly increased in non-African or in African ancestries in the admixed Central and Eastern Sahelian populations were verified for association with complex diseases in the Catalog of Published Genome-Wide Association Studies (downloaded from https://www.genome.gov/26525384 on the 2nd of April 2015) (Welter et al. 2014).

Analysis of Selection

We had to join some of the neighboring samples owing to low sample size, and for that we took into account language affiliation, subsistence system, and the results from population structure, so that reasonably homogeneous sets of populations could be obtained (as also performed by Pickrell et al. 2009, and having in mind the values indicated by them, of a minimum of ~40 chromosomes for iHS and as few as 20 chromosomes for XP-EHH). Thus, we joined: Gurmantche, Gurunsi, Mossi, and Songhai (recalled Burkina); Daza and Kanembu; Arabs and Nubians; and Turkana and Samburu. We estimated the iHS (Voight et al. 2006) by using the selscan package (Szpiech and Hernandez 2014). SNPs were prunned for minor allele frequency (MAF) >1% in each population. iHS tracks the decay of haplotype homozygosity for both the ancestral and derived haplotypes extending from a tested core SNP. Scores were calculated for each SNP within the population as a whole and standardized within each of 100 bins of allele frequency, using the norm tool. To facilitate comparison of genomic regions between populations and to gain power by detecting selective sweeps affecting the haplotypic structure at surrounding SNPs (Pickrell et al. 2009), we split the genome into non-overlapping segments of 100 kb, so that at least 20 SNPs are present in each segment. For each window, the proportion of absolute standardized iHS scores higher than 2 were calculated and used to order the windows (we confirmed that this measure was more robust in our data set than using simply the highest absolute iHS score, as basically the ordered genes were more similar between neighbor

populations when using that measure). Moreover, 100-kb windows were assigned a percentile score based on the proportion of extreme iHS values in the segment. In the 99% percentile range, we checked if significant windows could be collapsed owing to their tandem location, and we did so while maintaining the highest percentile value for the new collapsed window. Percentiles were recalculated for the remaining windows after collapsing. With this strategy, we diverged from others (Voight et al. 2006; Pickrell et al. 2009) because we confirmed that binning windows by their number of SNPs (in bins incremented by 10 each, and excluding windows with <20 SNPs) could decrease the significance of relevant regions; basically, the sharing of top candidate selected regions was higher between Sahelian populations in our strategy than when using bins.

The SelScan tool (Szpiech and Hernandez 2014) was also used to estimate XP-EHH. We calculated XP-EHH for the following pairs of populations: each African population compared with Italians, each Western African population compared with Oromo, and Eastern and Central Sahel populations compared with Gambians. Consistently with other tests for selection, only markers with MAF >1% were used for XP-EHH computation. The obtained XP-EHH values were normalized by subtracting genome-wide mean XP-EHH and dividing by standard deviation (Szpiech and Hernandez 2014). Whole genome was divided into 100-kb windows (same windows used in iHS analysis). In every window, top XP-EHH value was selected and if it falls into top 0.1% of XP-EHH values, the window was selected and was searched for genes in a whole 100-kb window and also in the 5-kb flanking region around the highest XP-EHH value. Previous reports (Pickrell et al. 2009) have shown that the maximum XP-EHH scores are more powerful as a statistic than the fraction of extreme scores, in opposition to what happens for iHS and we followed those indications in our analyses. Details are provided in supplementary fig. S2 (Supplementary Material 1 online).

Pathway Analysis – KEGG Pathway Database

Enrichments in every KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway (http://www.genome.jp/kegg/pathway.html, last accessed April 2, 2015) were tested for all genes identified as positively selected using the three following tests: iHS (genes in 100-kb window, top 300 windows), XP-EHH vs Italians (top 300 windows, only genes in the 5-kb flanking region around the top XP-EHH value were considered), and XP-EHH vs Gambians/Oromo (for Eastern and Central Sahel populations, we used Gambia as a reference population; for West Africans, we used Oromo; also, top 300 windows and genes in the 5-kb flanking region around the top XP-EHH). Every gene that appeared in one of the three selection tests received a score calculated from its percentile in the test (the higher the selection signal, the higher the score). If the gene was flagged in several tests, the individual scores were summed up for that gene within a pathway and then normalized to fit scale from 0 to 10. We produced heat maps for each pathway displaying the genes that appeared in the selection test in at least one population (Supplementary Material 3 online). We also produced a global heat map displaying all pathways, summing up the scores for the genes flagged as selected in that pathway (Supplementary Material 2 online). Details are provided in supplementary fig. S2 (Supplementary Material 1 online).

Results and Discussion

We first studied the genetic structure of Sahelian populations using ADMIXTURE (Alexander et al. 2009). Figure 1*C* highlights the strong impact of gene flow in this region for K = 7, which includes a Mozabite/North African component that allows to address the hypothesis of a North African ancestry in Fulani (other K plots and cross-validation are presented in supplementary figs. S3 and S4, Supplementary Material 1 online). Western African populations present varying proportions of two clusters, one being more frequent in Atlantic Western populations (orange in fig. 1*C*; 90% in Mandenka), whereas the other is more frequent in Western/Central populations, especially in Esan and Yoruba of Nigeria (dark blue; reaching 74–81% frequency). This main Western/Central component probably represents groups speaking Niger-Congo languages, including Narrow Bantu speakers, as can be verified by its high frequency (75%) in Luhya, a Bantu-speaking population from Kenya, who also presents a substantial proportion of Eastern African ancestry (23%). In clear contrast with Western Africans, Eastern Africans present a considerable input from Near Eastern (green color), Arabian (yellow color), and North African (light blue color) components (23–49%), with the exception of Turkana (7%), as well as some Somali, who instead show substantial fractions of Western/Central ancestry (27%), possibly received from the Bantu expansions in the region of the African Great Lakes. Both populations from Chad, the Daza and Kanembu, present a high Eastern African component (light gray; 41–61%), mixed with Atlantic Western (14–25%), Western/Central African (10–29%), and North African (5–13%; lower in the sedentary Kanembu and higher in semi-nomadic Daza) backgrounds. The nomadic Fulani present the reverse pattern for the Atlantic Western versus Eastern African components (55% and 11%, respectively), but the proportion of the North African component is even higher (23%) than in Daza (~10%). These results support the hypothesis of a North African origin and a Western to Central Africa past migration for Fulani. Notice that in ADMIXTURE results for K < 7, the Mozabite/North African component is not identified, being identical to the European component, indicating that these two components are very similar.

PCA of the data confirms these relationships between populations (fig. 1B and supplementary fig. S5, Supplementary

Material 1 online), as well as Wright's F_{ST} metrics (supplementary figs. S6 and S7, Supplementary Material 1 online). Altogether, our results support the role of the Sahelian Belt as a main corridor for human migrations across the African continent.

We next sought to identify genomic regions showing excess ancestry from non-Sahelian populations, as well as outliers for informative statistics on positive selection detection, and to describe the fine-scale geographical distribution of these selection signals across the Sahel. Figure 2 illustrates the top-10 iHS candidate selected regions observed in each Sahelian population, and signals of selection in Italians for comparison (supplementary table S11, Supplementary Material 1 online, reports the iHS significant regions in all populations). Western Sahelian samples share a higher amount of significant regions between them than the Eastern group, many of them in the 1% tail of the distribution. Given the North African influences in Fulani and the non-African (via Eastern African) influence in Daza + Kanembu, these Central Sahelian populations have a mixed pattern of selection of the observed in Western and Eastern Sahelian groups. The results for the top-10 XP-EHH, when comparing with Italians (supplementary fig. S16 and S17 and table S12, Supplementary Material 1 online), show a higher sharing of selected genes between Western and Eastern Sahelian pools than in iHS, compatible with the longer time needed for the stronger selective sweeps detected in XP-EHH analysis. When comparing Western with Eastern groups (supplementary fig. S14 and S15 and table S12, Supplementary Material 1 online), the pattern of top XP-EHH-based selected genes was very homogeneous within the Western group, as well as very homogeneous within the Eastern group, but different between them.

RFMix software (Maples et al. 2013) was used to deconvolute the genomes of admixed Sahelian populations into segments originating from their two main parental populations. Interestingly, some of the selection signals were associated with a local genomic enrichment of the non-predominant ancestry in the admixed Sahelian populations (supplementary tables S3–S8 and figs. S8–S13, Supplementary Material 1 online), as described below.

The best example of a widespread candidate selection signal across the Sahel is the malaria-associated *DARC* gene (fig. 3*F*), located on chromosome 1. This gene, the Duffy antigen/receptor for chemokines, encodes a membrane-bound chemokine receptor used by *Plasmodium vivax* for internalization into red blood cells (Demogines et al. 2012), so that Duffy-negative phenotype is thought to confer resistance against malaria caused by this parasite. There is complete fixation of the protective null allele in the Western region and, as we detected here, a local enrichment of African ancestry in the highly admixed Sudanese Arabs and Nubians (fig. 4B), where *P. vivax* is frequent. This evidence attests to the high selective pressure for *DARC* gene in the Sahel environmental context, even in its northeastern-most border. Although selection on the *DARC* gene has been identified a long time ago, our work properly contextualizes its importance in the Sahelian selective landscape.

Another long genomic candidate selected region across the Sahel Belt is placed on chromosome 17, but it shows distinct peaks in the iHS measure between Western and Eastern African regions. The iHS peak is located around the *ITGAE* gene in the Western populations, whereas it is centered on *TRPV1/SHPK* genes (78,300 bp apart) in the Central and Eastern samples (supplementary figs. S18 and S19, Supplementary Material 1 online). This can be an ongoing process of differentiating selection on a previously large selected genomic region. *ITGAE* is a receptor for E-cadherin and mediates adhesion of intra-epithelial T-lymphocytes to epithelial cell monolayers (involved in KEGG pathway regulation of actin cytoskeleton), playing a role in the immune system. Grossman et al. (2013) also detected signals of selection for this gene (non-synonymous SNP) in Yoruba, and further confirmed its functional impact through structural homology modeling and conservation analysis. The *TRPV1* gene encodes a receptor for capsaicin, the main pungent ingredient in hot chili peppers, which is also activated by noxious increases in temperature (Gavva et al. 2004), and *SHPK* controls glucose metabolism and acts as a modulator of macrophage activation (Haschemi et al. 2012).

Three other regions show signals of positive selection in almost all populations across Sahel, in the iHS analysis (fig. 2). One is located on chromosome 12 around the *SPATS2* gene (supplementary figs. S20 and S21, Supplementary Material 1 online), which plays a role in spermatogenesis (Senoo et al. 2002). Another is on chromosome 17, containing the keratin genes *KRT39* and *KRT40*, the latter being expressed during hair follicle differentiation (Langbein et al. 2007). On chromosome 4, a signal is detected on the *PIGG* gene, which is involved in the glycosylphosphatidylinositol anchor biosynthesis pathway, allowing the attachment of cell surface proteins to the cell membrane (Stokes et al. 2014).

Notwithstanding the ubiquitous selected genes across the Sahel, there were different profiles of candidate selected genes in the Western and Eastern sides. The main difference resides in strong signals of selection in several genes of the calcium and related heart and oxytocin pathways in the western region (fig. 3*A*–*C*), whereas, in the eastern side, glycerolipid and glycerophospholipid metabolism pathways displayed stronger signals of selection (fig. 3*D* and *E*). Oxytocin controls a wide variety of central and peripheral effects, especially the stimulation of uterine contractions during parturition and milk release in lactation, which are *per se* strong selection drivers, but it also influences cardiovascular regulation (Arrowsmith and Wray 2014), redounding effects of the tandem selected genes (*PIK3R1*, *CACNG3*, *RYR2*, and *KCNJ12*) in pathways related with heart, namely, on cardiac muscle contraction and adrenergic signaling in cardiomyocytes. By opposition, the Eastern trans-region selection signal displayed by the

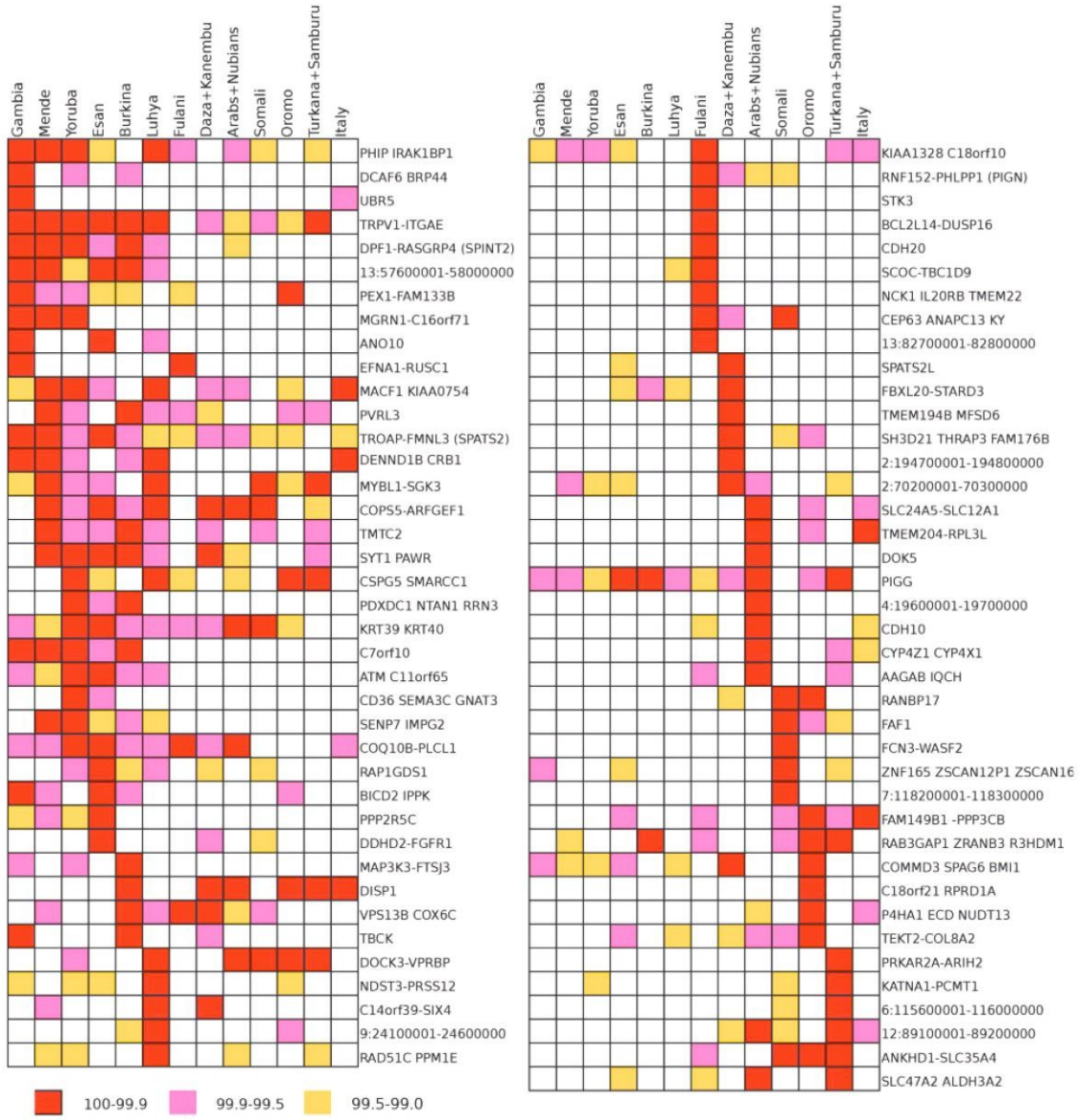

Fig. 2.—Top-10 iHS in each Sahelian population and matching selected genes in Italians. Some of the regions contain many genes, and only the first and last genes are indicated, with interesting genes reported inside brackets.

glycerolipid and glycerophospholipid metabolism pathways is mainly owing to the high values of selection in the *DGAT2* gene found on top XP-EHH vs West (supplementary fig. S15, Supplementary Material 1 online) and also the *DGKI* gene. This, together with the high value of selection detected in iHS in Eastern Sahel (except in Arab + Nubian) for the *RAB3GAP1* gene that is associated with cholesterol, testifies the importance of lipid metabolism in Eastern Sahel. It was already reported that the Eastern African pastoralist Maasai, whose diet of milk, blood, and meat is rich in lactose, fat, and cholesterol, display selection signals at the *RAB3GAP1/LCT/MCM6* region (Wagh et al. 2012). Here we show that this pattern is not only observed in nomads such as Maasai, Turkana, and Samburu, who may share a common/related origin, but also in the sedentary Oromo and Somali. Interestingly, that region in chromosome 2 is not

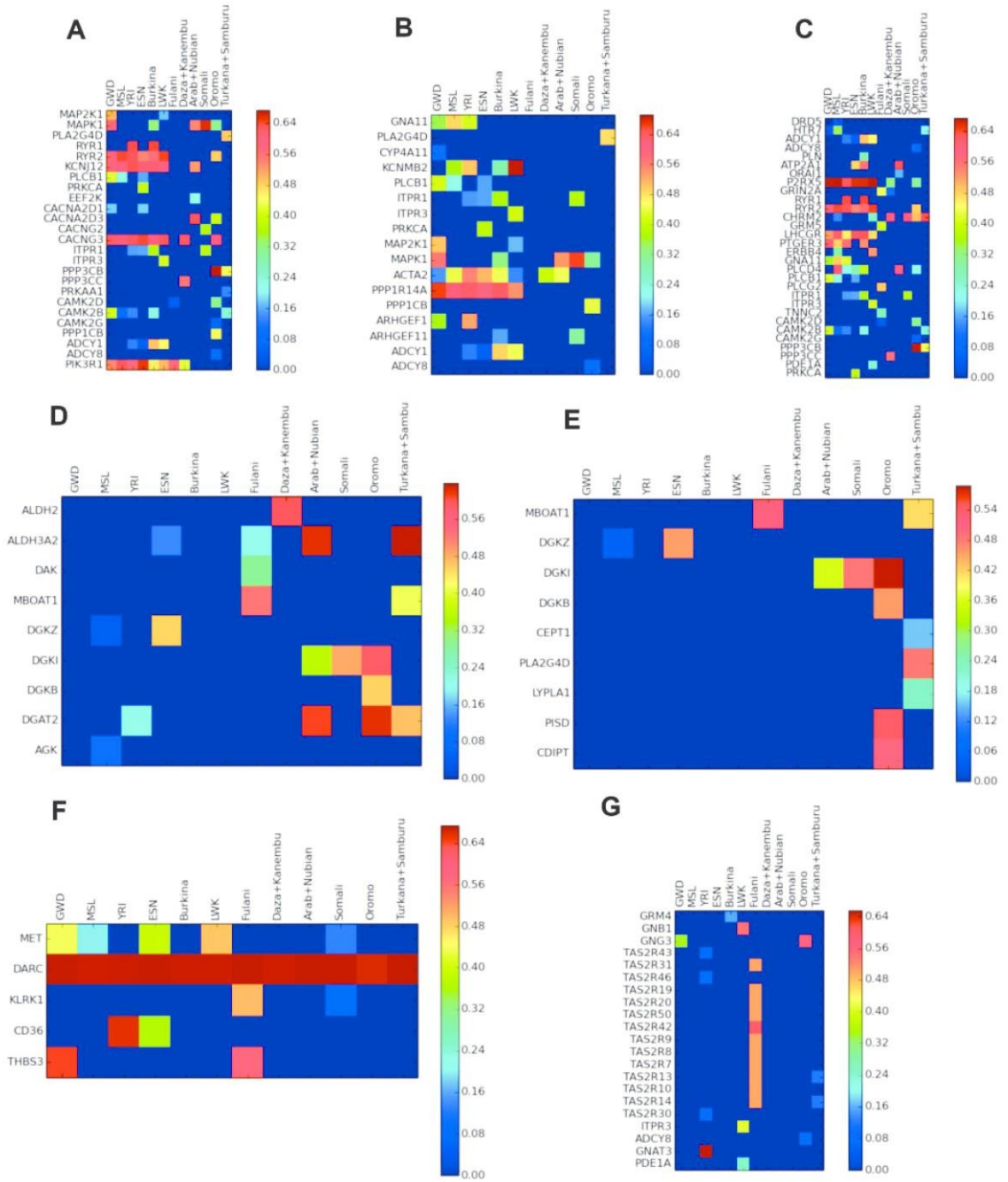

FIG. 3.—Selected genes in informative metabolic pathways (KEGG database). (*A*) Oxytocin signaling pathway. (*B*) Vascular smooth muscle contraction. (*C*) Calcium signaling pathway. (*D*) Glycerolipid metabolism. (*E*) Glycerophospholipid metabolism. (*F*) Malaria. (*G*) Taste transduction. GWD – Gambia; MSL – Mende; YRI – Yoruba; ESN – Esan; LWK – Luhya.

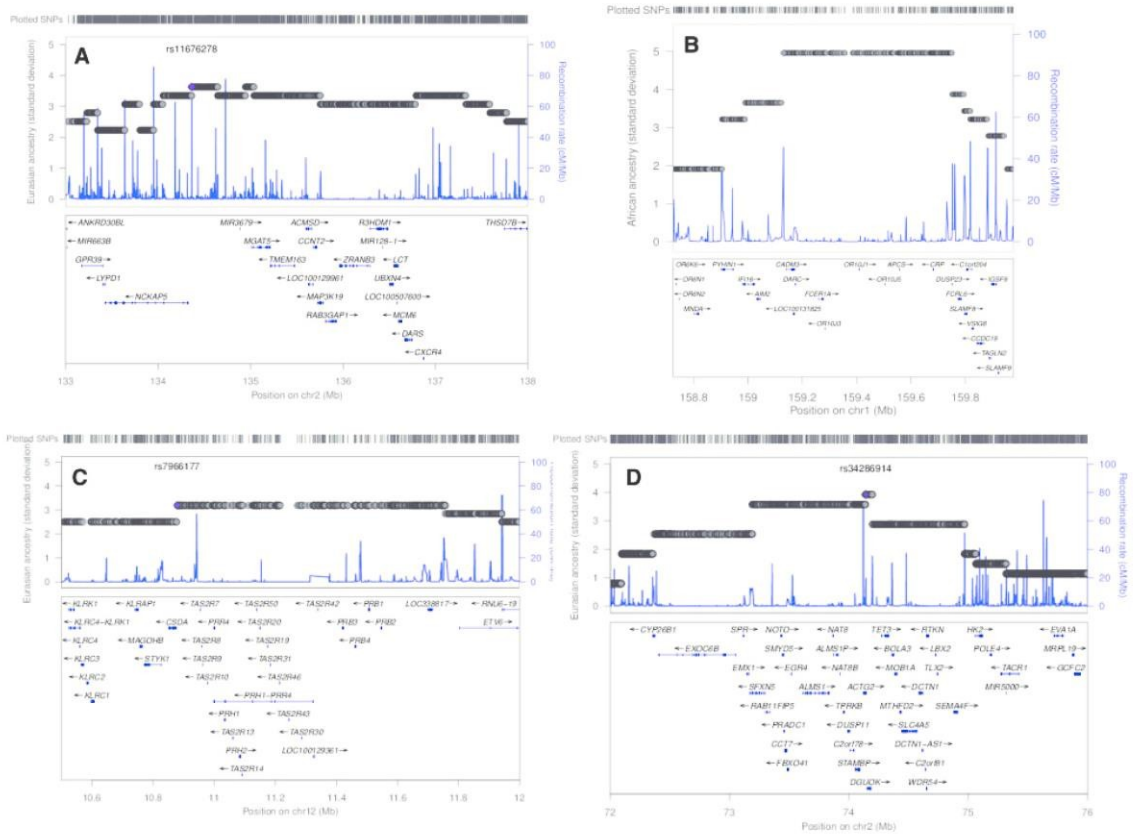

Fig. 4.—Locus zoom of a few enriched ancestry regions. (*A*) Chromosome 2 in Oromo. (*B*) Chromosome 1 in Sudanese Arabs + Nubians. (*C*) Chromosome 12 in Fulani. (*D*) Chromosome 2 in Turkana + Samburu.

autochthonous, having a significant non-African enrichment in Oromo (fig. 4*A*) that probably conferred an advantage to these populations, who have milk and blood as important food sources. Another non-African enriched region detected in Turkana + Samburu is located in another region of chromosome 2 (fig. 4*D*), containing the *ALMS1* gene associated with leprosy (Grossman et al. 2013) as well as the *NAT8* gene involved in creatinine levels and chronic kidney disease, which is frequent in African descendants. The Eurasian enrichment is probably a protection against chronic kidney disease in these Eastern groups.

The candidate selected regions restricted to Western Sahel have genes playing important roles. This is the case of a region in chromosome 19, rich in many genes, where the highest iHS (supplementary figs. S22 and S23, Supplementary Material 1 online) and XP-EHH vs East values are attained for the gene *SPINT2*, which has been associated with diarrhea (Heinz-Erian et al. 2009), and neighboring gene *CATSPERG*, which is required for sperm hyperactivated motility and male fertility (Wang et al. 2009). Other examples are: *BTRC*, previously associated with HIV (Nazari-Shafti et al. 2011); *TLR5*, suggested to alter NF-kB signaling in response to bacterial flagellin (Grossman et al. 2013); *PSMD8*, a member of the 26S proteasome involved in the regulation of transcription initiation (Durairaj and Kaiser 2014); and *LHCGR*, whose mutations result in disorders of male secondary sexual character development (Jeha et al. 2006).

The most geographically restricted selection signature was detected in Fulani, corresponding to a non-African ancestry enriched region in chromosome 12 (figs. 3*G* and 4*A*). This region contains *TAS2R* genes, which detect natural alkaloids such as quinine and strychnine (Kim et al. 2005; Reed et al. 2010; Ledda et al. 2014), possibly establishing a bridge with a Fulani ritual, of vital importance for some nomadic groups. The ceremony consists in a public flogging named "Sharo," a test of manhood: all youths must pass this ordeal without flinching to be considered as adults and eligible to get married (Adeola 2014). The courage is fortified by the previous drinking of a beverage (native beer or palm tree) containing seeds of the plant *Datura metel*, which initiates a stupefying,

narcotic effect. The seeds of this plant contain alkaloids (Oliver-Bever 1986) scopolamine or hyoscine, which depress the central nervous system, as well as hyoscyamine, which blocks all the body secretions, including the lachrymal glands, preventing tearing. Could non-African alleles in *TAS2R* genes on chromosome 12, introduced by the admixed origin of Fulani, allow a more efficient processing of these alkaloids contained in the beverage? This is an interesting hypothesis to be further tested, as it would be a striking example of sexual selection driving a significant excess of non-African ancestry. It must be reinforced that these *TAS2R* genes are distinct from the *TAS2R16* on chromosome 7, which detects salicin, a bitter β-glycoside anti-inflammatory compound, reported previously as having been under selection (through Fay, Wu's H, and McDonald–Kreitman tests) in Eurasian, East African, and Fulani populations (Li et al. 2011; Campbell et al. 2014). We did not detect signals of selection in *TAS2R16*.

Conclusion

In summary, by combining a rich population survey, a high-resolution genome-wide chip, and local ancestry and selection tools, we have demonstrated the power to dissect the palimpsest of complex interactions between cross and local selective pressures and demographic factors. We have confirmed independently selection signals described before, found new candidates to be further investigated, and provided a first glimpse of their spatial distribution across a considerable region of the African continent that has been playing a major role in human migrations along millennia. Signals are very strong in certain genes, persisting across Sahel, probably indicating that the selection event occurred in the ancestral African population, before the main Pleistocene migrations that established the bulk of the Sahelian genetic landscape (Soares et al. 2012; Rito et al. 2013). Malaria selection in the *DARC* gene is most probably an old (before 100,000 years ago) pathogen-driven selection force (Karlsson et al. 2014). But several differences exist between Western and Eastern regions, largely explained by the high admixture with non-African ancestry observed in the Eastern side and by recent pathogens that could have led to local adaptations. One possible example is the non-African enrichment signal detected in Eastern Sahel for *ALMS1* gene, which has been associated with leprosy based on GW association analysis in Indian patients (Grossman et al. 2013). Leprosy pathogen dates to around 12,000 years (Karlsson et al. 2014).

It is interesting that many of the enriched-ancestry and candidate selected regions are large and contain several genes that can contribute to adaptation to different selective factors. For instance, the Eastern Sahelian selected *RAB3GAP1/LCT/MCM6* region, related with lipid metabolism, also contains the *CXCR4*, determinant in HIV entrance into cells and tuberculosis resistance. The *TAS2R* region in Fulani contains *PRB3* gene, which acts as a bacterial receptor. The olfactory receptors *OR51B5* and *OR51B6* have been associated with sickle-cell anemia (Solovieff et al. 2010) but probably owing to their overlap with the *HbE* gene and neighboring *HbB* gene. An indirect evidence of their probable role in other non-olfactory sensory functions is their confirmed expression outside the olfactory epithelium, such as in kidney (Pluznick et al. 2009) and sperm (Spehr et al. 2003). *OR10J5* gene, located in the region containing *DARC* and other genes associated with hematology traits, has been said to play a role in angiogenesis and its expression in aorta and coronary artery has been confirmed (Kim et al. 2015).

The modeling of the African genetic diversity by selection driven by infectious diseases, since the origin of modern humans till present times, provides useful insights into natural ways of resistance. These natural strategies can potentially be mimicked pharmacologically, opening new avenues in the combat of infectious and other complex diseases. A good example of this potential is provided by the HIV resistance of *CCR5*-Δ32 homozygous, which is being used as the basis of stem cell transplantation trial therapies (Novembre and Han 2012).

Supplementary Material

Supplementary data S1–S3, figures S1–S23, and table S1–S12.24 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 290344 (EUROTAST). This project was also supported by the Grant Agency of the Czech Republic (13-37998S-P505). P.S. is supported by FCT (the Portuguese Foundation for Science and Technology), through FCT Investigator Programme (IF/01641/2013). IPATIMUP integrates the i3S Research Unit, which is partially supported by FCT. FEDER, COMPETE, and FCT fund IPATIMUP (PEst-C/SAU/LA0003/2013) and CBMA (PEst-OE/BIA/UI4050/2014).

Literature Cited

Abecasis GR, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.

Adeola BS. 2014. Datura Metel L: Analgesic or Hallucinogen? "Sharo" Perspective. Middle East J Sci Res. 21:993–997.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664.

Arrowsmith S, Wray S. 2014. Oxytocin: its mechanism of action and receptor signalling in the myometrium. J Neuroendocrinol. 26:356–369.

Botigue LR, et al. 2013. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proc Natl Acad Sci U S A. 110:11791–11796.

Bryc K, et al. 2010. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci U S A. 107:786–791.
Campbell MC, et al. 2014. Origin and differential selection of allelic variation at TAS2R16 associated with salicin bitter taste sensitivity in Africa. Mol Biol Evol. 31:288–302.
Campbell MC, Tishkoff SA. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet. 9:403–433.
Cerny V, et al. 2006. MtDNA of Fulani nomads and their genetic relationships to neighboring sedentary populations. Hum Biol. 78:9–27.
Cerny V, et al. 2009. Migration of Chadic speaking pastoralists within Africa based on population structure of Chad Basin and phylogeography of mitochondrial L3f haplogroup. BMC Evol Biol. 9:63.
Cerny V, et al. 2011. Genetic structure of pastoral and farmer populations in the African Sahel. Mol Biol Evol. 28:2491–2500.
Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156—2158.
Delaneau O, Marchini J, Zagury JF. 2012. A linear complexity phasing method for thousands of genomes. Nat Methods. 9:179–181.
Demogines A, Truong KA, Sawyer SL. 2012. Species-specific features of DARC, the primate receptor for Plasmodium vivax and Plasmodium knowlesi. Mol Biol Evol. 29:445–449.
Drake NA, Blench RM, Armitage SJ, Bristow CS, White KH. 2011. Ancient watercourses and biogeography of the Sahara explain the peopling of the desert. Proc Natl Acad Sci U S A. 108:458–462.
Durairaj G, Kaiser P. 2014. The 26S proteasome and initiation of gene transcription. Biomolecules 4:827–847.
Etefa T. 2012. Integration and peace in East Africa: a history of the Oromo Nation. New York: Palgrave Macmillan.
Gavva NR, et al. 2004. Molecular determinants of vanilloid sensitivity in TRPV1. J Biol Chem. 279:20283–20295.
Grossman SR, et al. 2013. Identifying recent adaptations in large-scale genomic data. Cell 152:703–713.
Gurdasani D, et al. 2015. The African Genome Variation Project shapes medical genetics in Africa. Nature 517:327–332.
Hanotte O, et al. 2002. African pastoralism: genetic imprints of origins and migrations. Science 296:336–339.
Haschemi A, et al. 2012. The sedoheptulose kinase CARKL directs macrophage polarization through control of glucose metabolism. Cell Metab. 15:813–826.
Heinz-Erian P, et al. 2009. Mutations in SPINT2 cause a syndromic form of congenital sodium diarrhea. Am J Hum Genet. 84:188–196.
Hernandez RD, et al. 2011. Classic selective sweeps were rare in recent human evolution. Science 331:920–924.
Hotez PJ, Kamath A. 2009. Neglected tropical diseases in sub-saharan Africa: review of their prevalence, distribution, and disease burden. PLoS Negl Trop Dis. 3:e412.
Huestis DL, Lehmann T. 2014. Ecophysiology of Anopheles gambiae s.l.: persistence in the Sahel. Infect Genet Evol. 28:648–661.
Jeha GS, Lowenthal ED, Chan WY, Wu SM, Karaviti LP. 2006. Variable presentation of precocious puberty associated with the D564G mutation of the LHCGR gene in children with testotoxicosis. J Pediatr. 149:271–274.
Karlsson EK, Kwiatkowski DP, Sabeti PC. 2014. Natural selection and infectious disease in human populations. Nat Rev Genet. 15:379–393.
Kim SH, et al. 2015. Expression of human olfactory receptor 10J5 in heart aorta, coronary artery, and endothelial cells and its functional role in angiogenesis. Biochem Biophys Res Commun. 460:404–408.
Kim U, Wooding S, Ricci D, Jorde LB, Drayna D. 2005. Worldwide haplotype diversity and coding sequence variation at human bitter taste receptor loci. Hum Mutat 26:199–204.
Lamphear J. 1988. The People of the Grey Bull: the origin and expansion of the Turkana. J Afr Hist. 29:27–39.
Langbein L, et al. 2007. Novel type I hair keratins K39 and K40 are the last to be expressed in differentiation of the hair: completion of the human hair keratin catalog. J Invest Dermatol. 127:1532–1535.
Ledda M, et al. 2014. GWAS of human bitter taste perception identifies new loci and reveals additional complexity of bitter taste genetics. Hum Mol Genet. 23:259–267.
Lewis IM. 1999. A pastoral democracy: a study of pastoralism and politics among the Northern Somali of the Horn of Africa. Oxford: James Currey Publishers.
Li JZ, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104.
Li H, Pakstis AJ, Kidd JR, Kidd KK. 2011. Selection on the human bitter taste gene, TAS2R16, in Eurasian populations. Hum Biol. 83:363–377.
Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. 93:278–288.
Marshall F, Hildebrand E. 2002. Cattle before crops: the Beginnings of Food Production in Africa. J World Prehistory 16:99–143.
Molesworth AM, et al. 2002. Where is the meningitis belt? Defining an area at risk of epidemic meningitis in Africa. Trans R Soc Trop Med Hyg. 96:242–249.
Nazari-Shafti TZ, et al. 2011. Mesenchymal stem cell derived hematopoietic cells are permissive to HIV-1 infection. Retrovirology 8:3.
Newman JL. 1995. The peopling of Africa: a geographic interpretation. New Haven: Yale University Press.
Novembre J, Han E. 2012. Human population structure and the adaptive response to pathogen-induced selection pressures. Philos Trans R Soc Lond B Biol. Sci. 367:878–886.
Oliver-Bever B. 1986. Medicinal plants in tropical West Africa. Cambridge: Cambridge University Press.
Patin E, et al. 2014. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nat Commun. 5:3163.
Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet. 2:e190.
Pereira L, et al. 2010. Linking the sub-Saharan and West Eurasian gene pools: maternal and paternal heritage of the Tuareg nomads from the African Sahel. Eur J Hum Genet. 18:915–923.
Pickrell JK, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 19:826–837.
Pickrell JK, et al. 2012. The genetic prehistory of southern Africa. Nat Commun. 3:1143.
Pickrell JK, et al. 2014. Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci U S A. 111:2632–2637.
Pluznick JL, et al. 2009. Functional expression of the olfactory signaling system in the kidney. Proc Natl Acad Sci U S A. 106:2059–2064.
Podgorna E, Soares P, Pereira L, Cerny V. 2013. The genetic impact of the lake chad basin population in North Africa as documented by mitochondrial diversity and internal variation of the L3e5 haplogroup. Ann Hum Genet. 77:513–523.
Prugnolle F, Manica A, Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. Curr Biol. 15:R159–160.
Purcell S, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81:559–575.
Reed DR, et al. 2010. The perception of quinine taste intensity is associated with common genetic variants in a bitter receptor cluster on chromosome 12. Hum Mol Genet. 19:4278–4285.
Rito T, et al. 2013. The first modern human dispersals across Africa. PLoS One 8:e80031.

Sabeti PC, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918.

Senoo M, Hoshino S, Mochida N, Matsumura Y, Habu S. 2002. Identification of a novel protein p59(scr), which is expressed at specific stages of mouse spermatogenesis. Biochem Biophys Res Commun. 292:992–998.

Silva M, et al. 2015. 60,000 years of interactions between Central and Eastern Africa documented by major African mitochondrial haplogroup L2. Sci Rep. 5:12526.

Smith MW, et al. 2004. A high-density admixture map for disease gene discovery in african americans. Am J Hum Genet. 74:1001–1013.

Soares P, et al. 2012. The expansion of mtDNA haplogroup L3 within and out of Africa. Mol Biol Evol. 29:915–927.

Solovieff N, et al. 2010. Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5′ olfactory receptor gene cluster. Blood 115:1815–1822.

Spehr M, et al. 2003. Identification of a testicular odorant receptor mediating human sperm chemotaxis. Science 299:2054–2058.

Spencer P. 1965. The Samburu: a study of gerontocracy in a nomadic tribe. Oxon: Routledge.

Steverding D. 2008. The history of African trypanosomiasis. Parasit Vectors. 1:3.

Stokes MJ, Murakami Y, Maeda Y, Kinoshita T, Morita YS. 2014. New insights into the functions of PIGF, a protein involved in the ethanolamine phosphate transfer steps of glycosylphosphatidylinositol biosynthesis. Biochem J. 463:249–256.

Szpiech ZA, Hernandez RD. 2014. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol. 31:2824–2827.

Teo YY, Small KS, Kwiatkowski DP. 2010. Methodological challenges of genome-wide association analysis in Africa. Nat Rev Genet. 11:149–160.

Tishkoff SA, et al. 2009. The genetic structure and history of Africans and African Americans. Science 324:1035–1044.

UNEP. 2008. AFRICA Atlas of our changing environment. Sioux Falls (SD): United Nations Environment Programme.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:e72.

Wagh K, et al. 2012. Lactase persistence and lipid pathway selection in the Maasai. PLoS One 7:e44751.

Wang H, Liu J, Cho KH, Ren D. 2009. A novel, single, transmembrane protein CATSPERG is associated with CATSPER1 channel protein. Biol Reprod. 81:539–544.

Welter D, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42:D1001–1006.

WHO. 2012. Sahel food and health crisis: emergency health strategy. West Africa Regional Health Working Group. p. 28.

Wolfe ND, Dunavan CP, Diamond J. 2007. Origins of major human infectious diseases. Nature 447:279–283.

Zakharia F, et al. 2009. Characterizing the admixed African ancestry of African Americans. Genome Biol. 10:R141.

Associate editor: Naruya Saitou

3.2 Paper II

OSBPL10, *RXRA* and lipid metabolism confer African-ancestry protection against haemorrhagic fever in admixed Cubans.

In preparation.

OSBPL10, *RXRA* and lipid metabolism confer African-ancestry protection against dengue haemorrhagic fever in admixed Cubans

Beatriz Sierra[1±*], Petr Triska[2,3,4±], Pedro Soares[3], Gissel Garcia[1], Ana B. Perez[1], Eglys Aguirre[1], Marisa Oliveira[2,3,4,5], Bruno Cavadas[2,3], Béatrice Regnault[5], Mayling Alvarez[1], Didye Ruiz[1], David C. Samuels[6], Anavaj Sakuntabhai[7], Luisa Pereira[2,3,8*], Maria G. Guzman[1]

[1] Virology Department, PAHO/WHO Collaborating Center for the Study of Dengue and its Vector, Pedro Kourí Institute of Tropical Medicine (IPK), 601 Havana, Cuba
[2] Instituto de Investigação e Inovação em Saúde (i3S), 4200-135 Porto, Portugal
[3] Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), 4200-135 Porto, Portugal
[4] Instituto de Ciências Biomédicas Abel Salazar (ICBAS), Universidade do Porto, 4050-313 Porto, Portugal
[5] Eukaryote Genotyping Platform, Genopole Pasteur Institute, 75724 Paris, France
[6] Vanderbilt Genetics Institute, Department of Molecular Physiology and Biophysics, Vanderbilt University School of Medicine, Nashville, 37232-0700 TN, USA
[7] Functional Genetics of Infectious Diseases Unit, Pasteur Institute, 75015 Paris, France
[8] Faculdade de Medicina da Universidade do Porto (FMUP), 4200-319 Porto, Portugal

Running title: Admixture mapping in Cuban individuals infected by dengue
Keywords: Dengue; Cuba; GWAS; admixture mapping; infectious diseases; lipid metabolism

[±] Authors contributed equally to the present work
* Correspondence should be addressed to Beatriz Sierra (email: siebet@ipk.sld.cu) or Luisa Pereira (email: lpereira@ipatimup.pt).

Abbreviations: DF - dengue fever; DENV – dengue virus; DSS - dengue shock syndrome; DHF - dengue haemorrhagic fever; HH – Havana haemorrhagic; HF – Havana fever; HA - Havana asymptomatic; HC – Havana control; GH – Guantanamo haemorrhagic; GF – Guantanamo fever; GA - Guantanamo asymptomatic; GC – Guantanamo control; HCG - haemorrhagic comparison group; FCG - fever comparison group; OCG - overall comparison group; T- triglyceride; LDL - low-density lipoproteins; HDL - high high-density lipoproteins

Abstract

A 2.5 million SNP GWAS in a dengue fever cohort from the admixed Cuban population confirmed African protection against the haemorrhagic phenotype. Admixture mapping and fine-matched association test identified the candidate genes *OSBPL10* and *RXRA*, with most significant SNPs outside the coding region in segments of inferred weak enhancers, promoters and lncRNAs. *OSBPL10* presents a significantly lower expression in Africans compared with Europeans, while for *RXRA* several SNPs in regulatory regions may regulate differentially its transcription between Africans and Europeans. The expression of these genes changes along dengue disease progression in Cuban patients. Also, significantly altered *OSBPL10* and *RXRA* expressions were confirmed in the enlarged LXR/RXR pathway (integrating immune functions and lipid metabolism) in macrophages, by gene set enrichment analysis of Thai dengue transcriptome data. Our genomic evidence of protection against dengue haemorrhagic fever, conferred by *OSBPL10*, *RXRA* and related lipid metabolism, point out potential therapeutic applications.

Dengue is an emerging arthropod-born viral disease caused by infection with any of the four dengue viruses (DENV-1 to 4), grouped in a complex of the *Flaviviridae* family and genus *Flavivirus*. The virus is transmitted between humans by *Aedes aegypti* and *Aedes albopictus* mosquitoes. Morbidity and mortality associated with the severe dengue infection render this disease a major increasing public health problem throughout tropical and subtropical regions of the world, and is attracting awareness in Europe and the United States as climate change and globalisation enlarge the geographic dispersion of the virus and vector[1]. A wide spectrum of disease manifestations is seen, ranging from subclinical infection, a relatively mild, self-limited infection known as dengue fever (DF), to severe disease, dengue haemorrhagic fever (DHF), characterized by an increase in vascular permeability, frequently accompanied with thrombocytopenia and haemostatic dysfunction and severe bleeding, which may evolve to a life-threatening hypovolemic shock (dengue shock syndrome, DSS)[2]. But only a small proportion of antibody-positive individuals develops DHF/DSS, the vast majority suffering asymptomatic infection or mild disease[3]. This differential susceptibility to disease severity indicates that host genetic factors may influence DENV disease outcome, acting in a complex interplay with viral and environmental factors. Diverse single nucleotide polymorphisms (SNPs) in genes such as *HLA-I, HLA-II*, *TNF-α*, *IL-10*, *TGF-β1*, *FcγRIIa*, *VDR*, *CD209* and *OAS* have been associated with symptomatic dengue or have been considered protective against the disease, in Asian and Latin American populations[4]. Also *MICA* and *MICB* genes were associated with susceptibility to dengue in Cuba[5], partially overlapping results reported in the only genome-wide association study (GWAS) performed so far[6]. That study was done in Vietnamese children and showed significant association of polymorphisms within *MICB* and *PLCE1* genes with DSS[6].

Evidence supporting the impact of human genetic factors on dengue infection outcome also comes from the trend observed in certain populations and ethnic groups to develop severe DHF/DSS symptoms, while others only develop milder disease or asymptomatic infection[7]. As early as 1906, it was reported that Cuban dark-skinned individuals showed a remarkable resistance against dengue disease compared with light-skinned individuals[8]. However, it was during the 1981 Cuban DHF/DSS epidemic of DENV-2 that ethnicity was recognized for the first time as a possible host risk factor, with highest severity in light-skinned individuals. This observation has been confirmed in several dengue outbreaks in Cuba[9]. The low occurrence of dengue disease in

autochthonous Haitians[10], as well as in African populations[11] add further evidences to the role of ethnicity in dengue disease outcome.

The genome-wide era allows us to assess this hypothesis of ancestry-related susceptibility to dengue illness. A first study was applied in the Colombian population[12], through a chip containing 30 ancestry informative markers, which confirmed the protective effect of African ancestry against severe dengue outcomes (odds ratios, ORs in 0.963-0.971 interval). A genome-wide replication study, based on thousands of SNPs, is still needed to be performed in other Latin American admixed populations, as it would allow the identification in an unbiased way of the most-significant African protective genes against dengue disease.

Till recently, population structure as occurs in admixed populations was a major confounding factor, requiring strategies for correction of the association p-values[13]. But admixed populations are a great advantage in cases of differential ancestry-conferred susceptibility/resistance to a disease through the use of admixture mapping[14]. The rationale of admixture mapping is that the ancestry blocks will be distributed at random across the genome, reflecting the admixture proportions of the parental ancestries, except in candidate gene locations where statistically significantly different proportions for the ancestry with higher disease levels will be observed in cases versus controls. It has been shown that this test is statistically more powerful than traditional GWAS[15]: around 250 samples can provide a 60% power to detect a two-fold risk due to ancestry, compared to the thousands of samples required in GWAS. In fact, because of the recentness of admixture, the typical ancestry blocks are significantly larger than haplotype blocks, thus lowering the multiple testing burden. However, an adverse aspect of this is that admixture mapping peaks cover hundreds of kilobases, rendering it difficult to identify the causal variant. This strategy has been successfully applied in African-Americans and Latin-Americans, in association with various diseases, such as asthma[16] and type 2 diabetes[17].

Cuba is extremely advantageous for genetic studies of DHF susceptibility. The current Cuban population is mainly derived from the mix of two well-defined populations: European colonizers, who began to arrive in 1492 from diverse areas of the Iberian Peninsula; and African slaves, arriving in the 16th century, mainly from West Africa. The contribution of the first aboriginal Cuban inhabitants, almost totally exterminated during the Spanish conquest, is almost negligible[18]. It may also be easier to identify genetic and non-genetic determinants, as the insular population has been exposed to

identical previous dengue infections[9]. In addition, the Cuban health system centralized a multidisciplinary management program of dengue epidemics, using standardized clinical criteria guidelines for dengue prevention and control[9]. This program also facilitates the detection of asymptomatic cases, which are usually overlooked, despite being the best control group. Consequently, we conducted a GWAS of 2.5 million SNPs in 274 Cubans, including patients (DF and DHF) of the 2006 dengue epidemic, from Havana (west) and Guantanamo (east) cities, and geographically matched asymptomatic individuals and population controls. The high level of African and non-African population admixture in Cuba[19] enabled us to apply the first admixture mapping, thus facilitating the identification of candidate markers ethnically associated with dengue infection.

Results

Global ancestry influence in dengue infection outcome. The genome-wide Cuban screening confirmed that all individuals in this study have a mixed ancestral composition. The main ancestry backgrounds derive from Africa and Europe, but the range follows the entire spectrum of admixture, from nearly 0% African and 90% European to the inverse ratio (Fig. 1A for K=4), while the remaining 10% are from Native American and East Asian influences. Comparing the population control groups from Havana and Guantanamo (HC and GC) as references for the two geographical regions, the average proportions of the African component are statistically different (25.2% and 35.3%, respectively; two-tailed Wilcoxon rank-sum test $p=1.43x10^{-3}$), identical to published values [19]. The Native American component was 6.5% in Havana and 13.5% in Guantanamo, values statistically significantly different ($p=1x10^{-6}$); while the East Asian component was of 1.9% and 0.7% respectively (p=0.224). A finer description of Cuban ancestry is presented in Supplementary section 1.2 (Supplementary Fig. 1-2). This includes confirmation that, despite the statistical differences in the African/European components between Havana and Guantanamo, the sub-structures within those two components in the two cities are identical, not favouring differential migration events into the two parts of the island (Supplementary Fig. 3-4).

The analysis of the global ancestry patterns among the Cuban cohorts allows us to evaluate the ancestry influence in the susceptibility to dengue infection. As the African, European, Native American and East Asian components sum up necessarily to 100%, enforcing co-linearity between the variables, and the European and African components are highly negatively correlated ($r^2=0.9319$; Supplementary Fig. 5), we discarded the European component from the analysis. East Asian ancestry was also ignored as it is negligible (<3%). The average African ancestry (Fig. 1B) is significantly lower in DHF (22.9%) when compared with DF, controls and especially asymptomatic groups (30.6%, p-value=0.025; 30.0%, p-value=0.041; 34.7%, p-value=0.013, respectively). Therefore, these results confirmed that African ancestry is protective against the DHF phenotype, and the obtained odds ratios are very similar to the ones reported in Colombia[12] (Table 1). Nevertheless, the evidence of ancestry influence in dengue is not so straightforward when the samples are divided according to the city of origin. In Havana, the proportion of African ancestry is even more significantly lower in DHF (10.3%) compared with DF, controls and especially asymptomatic (24.4%, p-value=0.009; 25.2%, p-

value=0.015; 33.0%, p-value=0.002, respectively); while in Guantanamo, the African averages are statistically identical between all groups (35.3% in controls, 38.1% asymptomatic, 36.4% DHF and 36.0% DF). The odds ratios of African ancestry protection in Havana are even more protective than in Colombia (Table 1). For the Native American component, there were no statistical differences between Cuban groups (Supplementary Table 3).

By applying an iterative model, we further tested if African and Native American ancestries, as well as the location inside the island, should be taken as important variables in the dengue extreme phenotypes (asymptomatic and DHF). A significant p-value was observed for the protection conferred by African ancestry against DHF in Havana (p-value=0.002; Fig. 1C), while the Native American protection in Havana was not statistically significant and no significant ancestry influence was observed in Guantanamo.

Our data shows that there is an African protection conferred against haemorrhagic dengue phenotype in Cuba, but other currently unknown confounding factors render it a complex relation, even in such a geographical restricted scenario as Cuba.

Fine-matched corrected population structure followed by association evaluation. As the relative sub-African proportions within the African component were identical in Havana and Guantanamo samples (Supplementary Fig. 3), we grouped individuals by class of dengue phenotypes, independently of their origin. We applied a fine-matched population structure correction, based on the global African ancestry, as described in Methods. This is not so extreme as the traditional association test corrected for population structure by PCA information[13]. We then conducted association tests in the comparison groups (Fig. 2A; Supplementary Fig. 6 and Supplementary Tables 4-6).

The highest significant p-values were detected in the DHF comparison (HCG) for six consecutive SNPs in chromosome 3, in a region containing the *OSBPL10* (oxysterol binding protein-like 10) gene. This gene product, coded by the reverse DNA strand, is involved in lipid transport and steroid metabolism. The six SNPs extend for 8,370 bps, are highly linked (Supplementary Fig. 10-12), and half of them are located in the overlapping region with the *ZNF860* (zinc finger protein 860) gene (Fig. 2B), which is coded by the forward DNA strand. Curiously, the most frequent haplotypes in African and European populations consist totally of the alternative alleles for all six SNPs, and attain frequencies higher than 50% in those respective continents (Fig. 2C). The odds

ratio calculated in DHF is 0.25 [95% CI 0.13-0.47] for the African haplotype (Table 2). We will analyse this gene in detail in another section.

Several other SNPs located on genes (mostly one or two SNPs per gene) reached association p-values of 10^{-5} in the three comparison groups. Using information from the 1000 Genomes database, it was possible to confirm the ancestry of the protective allele in these SNPs: some had a clearly higher African frequency; others were more frequent in Europeans, Asians or both; a few had similar allele frequency across the three population groups; and some genes had both African and non-African protective alleles (such as *CAMK1D* and *PIK3AP1*). We checked which of these genes would have differential gene expression between dengue patients and control/convalescent subjects in the whole blood transcriptome obtained in a Thai dengue dataset[20], as an additional indication of its potential functional implication in dengue illness. In the African cohort (Supplementary Fig. 7), besides *OSBPL10*, a few kinases or kinase-related genes are amongst the most significant genes (functional information from GeneCards® Human Gene Database or otherwise referred): *CAMK1D* (calcium/calmodulin-dependent protein kinase ID) associated with chemokine, thrombin signalling and xenobiotic metabolism (through RXRA-CAR dimers), activation of neutrophil cells and apoptosis of erythroleukemia cells; *MAPKAPK5* (mitogen-activated protein kinase-activated protein kinase 5), that responds to cellular stress and pro-inflammatory cytokines, and the use of a specific inhibitor of this gene blocked DENV assembly[21]; *PIK3AP1* (phosphoinositide-3-kinase adaptor protein 1) which may interact with *OSBPL10*, as it is possible that this gene is activated by phosphatidylinositol-3-phosphate (PI3P)[22], and links Toll-like receptor signaling to PI3K activation preventing excessive inflammatory cytokine production; *SNRK* (sucrose nonfermenting related kinase) possibly expressed in the liver secretome; *GNA14* (guanine nucleotide binding protein (G protein), alpha 14), activates PLC (detected in the Vietnamese dengue GWAS) protein that generates diacylglycerol which further activates *PPARA* to forms heterodimers with *RXRA*, positively controlling the expression of genes related with lipid metabolism[23]; DAB-1 (dab, reelin signal transducer, homolog 1 (*Drosophila*)) may play a role in PI3K binding.

The non-African set (Supplementary Fig. 8) has more variable and generalized functions, dealing with hemostasis (*DOCK10*), phospholipid binding (*PLEKHM1L*), protein serine/threonine kinase activity and ribosomal protein S6 kinase activity (*RPS6KA2*), protein homodimerization activity and HMG box domain binding (*OLIG2*),

sequence-specific DNA binding transcription factor activity and chromatin binding (*TSHZ3*).

Fine-matched corrected population structure followed by admixture mapping. We implemented admixture mapping, by using the RFMix algorithm in the comparison groups (Supplementary Fig. 9-11 and Supplementary Tables 7-9). Two of the regions displaying a significantly higher proportion of African ancestry in the asymptomatic/control subjects overlap in DHF and DF comparison groups (HCG and FCG). One is on chromosome 9, containing the *RXRA* (retinoid X receptor alpha), *COL5A1* (collagen type V alpha 1) and *FCN2* (ficolin (collagen/fibrinogen domain containing lectin) 2) genes. The region *RXRA-COL5A1* had already been identified in a GWAS in Latinos, associated with central corneal thickness[24] and the identified SNP was located in a long non-coding RNA (lncRNA) placed between the two genes. We confirmed from the expression data on dengue[20] that only *RXRA* has statistically significant altered expression (Supplementary Fig. 12). There are several lines of evidence linking retinoid receptors with infectious diseases and dengue [25], leading us to explore this gene in more detail in the next section.

The other overlapping region is on chromosome 7, containing the *PTPRN2* (protein tyrosine phosphatase, receptor type, N polypeptide 2), *NCAPG2* (non-SMC condensin II complex, subunit G2), *ESYT2* (extended synaptotagmin-like protein 2), *WDR60* (WD repeat domain 60) and *VIPR2* (vasoactive intestinal peptide receptor 2) genes. Of these, *PTPRN2*, *NCAPG2* and *ESYT2* have significantly differentiated expressions in the Thai dengue dataset[20]. *PTPRN2* is an interesting finding as it may dephosphorylate PI3P[26], but it is surpassed in terms of the top significantly associated gene in this region by *VIPR2* (Supplementary Table 12), which is not differently expressed in Thai dengue patients. This gene was also detected as being under selection in the DHF group in the XP-EHH analysis (Supplementary Table 15), and encodes a receptor for the small neuropeptide vasoactive intestinal peptide, involved in water and ion flux.

In the overall comparison group (OCG), this region 7 is also detected (Supplementary Table 9), as well as a long region in chromosome 22 overlapping with the FCG results (Supplementary Table 8). Interestingly, the *OSBP2* gene is in this region, being up-regulated in controls/convalescents, while the chromosome 3 located *OSBPL10* is up-regulated in patients. Another interesting gene in chromosome 22 is *SEC14L2* (and other family members), which has recently been shown to be essential for hepatitis C

virus (HCV) replication in cell culture[27]. No differential expression for the *SEC14L2* gene was reported in the Thai dataset. Nevertheless, the top p-values (Supplementary Table 13) in this region are for genes *SYN3* (synapsin III; one SNP detected also in the association test in FCG) and *TTC28* (tetratricopeptide repeat domain 28), with no evident link to viral diseases in the literature so far.

Exploring in depth *OSBPL10* and *RXRA* genes. For *OSBPL10*, we began by checking the expression impact of the two alternative ancestry haplotypes by using the transcriptome dataset from the 1000 Genomes project lymphoblastoid cell lines[28]. The homozygous individuals for the African haplotype have a significantly reduced (by half) expression when compared with the individuals homozygous for the European haplotype (two-tailed Wilcoxon rank-sum test $p<0.001$; mean African homozygous = 0.145; mean European homozygous = 0.257; Fig. 2D).

Given that the haplotype SNPs are non-protein coding, we then checked if they are located in regulatory regions. The tool to infer promoters[29] showed that there are two promoters for *OSBPL10* and one for *ZNF860* (Supplementary Fig. 17), but that the haplotype region is 2.5 kb from the OSBPL10_1 and ZNF860 promoters (OSBPL10_2 is even farther). This tool also infers three HMR conserved transcription factor binding sites within the beginning of the haplotype region, one being recognised by *STAT5A* (signal transducer and activator of transcription 5A), which is activated by a number of cytokine and growth hormone receptors, playing a key role in the transformation of B and T lymphocytes[30]. The haplotype region can also be related with a few weaker enhancer regions (Supplementary Table 17). Two haplotype SNPs (rs4600849 and rs11129475) are located in weak enhancers detected in different cell types, and two contiguous SNPs not in the chip (rs35108900 and rs57434781) are located in a weak enhancer detected in the hepatic HepG2 cell line. Several SNPs located in the *OSBPL10* haplotype are recognised by many transcription factors, including *STAT* and *RXRA*.

Besides the *OSBPL10* haplotype, the segment immediately 5' to it (within the gene) has some SNPs with significant p-values in the association test (Supplementary Table 10), that are regulatory regions in several cell types (Supplementary Fig. 18). According to the HGDP selection browser[31], this extended *OSBPL10* segment (haplotype and immediate 5' region) has been under positive selection in the African population of Yoruba and also (albeit less strongly) in Asia (iHS measure; Supplementary Fig. 19).

We applied another positive selection measure, XP-EHH (Supplementary Fig. 13; Supplementary Table 15), which detects stronger selective sweeps, to the Cuban HCG, and verified that the *OSBPL10* gene was under positive selection. Therefore, it seems that the *OSBPL10* haplotype and its 5' region, although non-coding, bear several expression regulatory regions containing SNPs with large differences in allele frequencies between Africans and Europeans, due to strong selection events.

In relation to *RXRA*, we checked the SNPs with significant association p-values in the *RXRA-COL5A1* region in the Cuban HCG (Fig. 3 and Supplementary Table 11), and verified that two locations have higher significant p-values ($p<0.01$) and OR between 0.103-0.436 (Table 2). The most significant p-values are for three intergenic SNPs (rs4262378, rs4424343 and rs3118593) located in inferred enhancers (Supplementary Table 17). The other location is immediately before and in the beginning of *RXRA*, and contains SNPs rs12339163 and rs62576287 (the latter only exists in Africa with a 9% frequency) placed in poised and weak promoters and weak enhancers. Additionally, two other close intergenic SNPs attain significant p-values just above the 0.01 threshold, are polymorphic only in African populations (10% for rs76917123-allele G and 4% for rs114989133-allele G) and are located in inferred enhancers. From these various significant SNPs, only rs3118593 is within a promoter and a lncRNA (*RP11-473E2.4*; Fig. 3), and interestingly, according to the GTEx portal, the patterns of expression for *RXRA* and *RP11-473E2.4* genes in the various human tissues are totally opposite (Supplementary Fig. 20), with *RXRA* being mainly expressed in liver, muscle, skin and whole blood while almost not expressed in brain and testis, and the other way around for *RP11-473E2.4*. Three other lncRNAs surround the *RXRA* gene, but are expressed in just one or two tissues (Supplementary Fig. 21-22) and at low levels. Given the opposite expression pattern between the *RXRA* and *RP11-473E2.4* genes, and the fact that rs3118593 allele A has been found to be a regulatory region variant in ten cell lines (according to Ensembl), we checked whether this SNP could confer differential *RXRA* expression. The genotypes for the rs3118593 SNP have opposite frequencies in Africa and Europe (0.436 AA, 0.460 AC and 0.104 CC; 0.085 AA, 0.425 AC and 0.489 CC, respectively; according to 1000 Genomes Project), but the *RXRA* mean expression is very similar in all genotypes (Supplementary Fig. 23), being slightly lower in Africans than in Europeans (except for CC). This result indicates that this SNP must not be the only control of *RXRA* expression; so we performed further association tests for the whole *RXRA-COL5A1* region in the 1000 Genomes transcriptome data. First, we

compared Africans having low (n=6; lower than 10 RPKM) and high (n=8; higher than 20 RPKM) *RXRA* expression; and second, we did the same comparison in Europeans (n=42 and n=39, respectively). Results show that several SNPs surrounding *RXRA* gene can be regions for its expression regulation (Fig. 3; Tables S18-S19), and these are more frequent in Africans than in Europeans, including in the *RP11-473E2.4* lncRNA gene. A reasonable hypothesis would be that the advantage in Africans is related with a faster control of *RXRA* expression.

Table 2 also sums up the tests were statistical significant evidence was detected for *OSBPL10* and *RXRA* genes.

***OSBPL10* and *RXRA* expression in dengue patients and focused enrichment analysis.** We checked the expression of *RXRA* and *OSBPL10* in Cuban patients along the infection process (Fig. 4). The mRNA expression of *RXRA* was significantly higher during convalescence (day 30) compared to day 3 ($p=0.027$), and also higher than at day 7 after fever onset (not significant). There were no significant differences between days 3 and 7 after fever onset. These results are comparable with the ones for the Thai dataset, where *RXRA* expression is significantly decreased in DF and DHF cohorts when compared both with controls and convalescent, indicating that this gene expression is decreased along the disease course, and only returns to normal values in convalescence.

The mRNA expression of *OSBPL10* was significantly increased by day 7 and in convalescence, when compared with day 3 ($p<0.001$ for both). There is a decrease between day 7 and convalescence, but it is not significant. In Thai, the DHF group has a significantly higher *OSBPL10* expression compared with all other groups. Considering both datasets, it seems that *OSBPL10* expression is very low in the acute phase, increases significantly at the end of the acute phase, and decreases again in convalescence. For both genes, results in Cuban patients showing warning signs (as described in Methods section) were similar to the ones observed in the totality of patients.

We also performed a focused gene enrichment analysis for the LXR/RXR interaction pathway, related to cholesterol metabolism and cytokine production in macrophages, where *RXRA* and *OSBPL10* play a central role (Fig. 5). Again, we took advantage of the Thai transcriptome dataset in whole blood as a surrogate of LXR/RXR interaction

pathway in macrophages. Unfortunately, there are no robust transcriptome dataset for hepatocytes, in which the same pathway occurs without the cytokine production.

We split the pathway into three gene sets, lipid metabolism, LXR/RXR activation and NF-kB activation (Supplementary Table 20), and used GSEA[32] to assess the statistical significance of their enrichment score. Results for comparisons of DHF and DF versus convalescent are represented in Fig. 6 (comparisons of DHF and DF versus controls are shown in Supplementary Fig. 24). The lipid metabolism set of genes is always significantly upregulated in patients, and the main contributing upregulated genes are *OSBPL10*, *LDLR* and *MSR1* (together with *ABCA1*, *CD36*, *SREBF2* and *INSIG2* in DF). *LDLR*, *CD36* and *MSR1* mediate the entrance of LDL (low-density lipoprotein) to the cell. *ABCA1* is one of the cholesterol efflux pumps. *SREBF2* and *INSIG2* activate LXR/RXR dimers in situations of high cholesterol content. DF patients seem to have more lipid-controlling genes upregulated than do DHF, including some feed-back control to decrease the high cholesterol content. *NF-kB* expression is always upregulated in the convalescents (significantly in DHF comparison), or, biologically more meaningful, down-regulated in patients. The genes contributing to this are both nuclear transcription controllers *RXRA* and *NF-kB* (*REL*, *RELB* and *NF-kB2*), the *NF-kB* controlled *COX2/PTGS2* and *IL1-β* genes, and the membrane receptors *TNFR*, *IL1R*, *LY96*, *TLR4*. The LXR/RXR activation set of genes is never statistically significant, but the most contributing genes are *NR1H3/LXRA*, *ABCA1*, *ARG2* and *NCOR1* up-regulated in DF, and *ABCG1*, *LPL* and *RXRA* down-regulated in DHF.

This analysis reinforces the importance of the enlarged pathway for LXR/RXR activation, including the cholesterol/lipids metabolism and the NF-kB control of cytokines in dengue disease. We demonstrated that the African protective genes *OSBPL10* and *RXRA*, which play central roles in this pathway, are always identified as top differentially expressed genes. We made another test by including the 12 *OSBP* family members in the analysis and confirmed that the *OSBPL10* gene is the most significantly overexpressed in patients (Supplementary Fig. 25).

Discussion

Since the out-of-Africa migration at around 60,000 years ago[33], human populations have been structured in three main groups, African, European and Asian. The independent selective pressures acting upon these groups can lead different genes to be selected in adaptation to the same pathogen or group of related agents. The diverse resistance genes can also occur in a common crucial pathway or in different pathways of additive importance to the disease process.

The present study presents evidence that supports the higher African-ancestry-conferred resistance to DHF, in comparison with European background, concurring with the Chacon-Duque et al. report[12]. It is unlikely that DENV has exerted this selective pressure, since its associated mortality is low and it does not alter reproduction. It has been discussed that the selective pressure conferred by another African-originated flavivirus, the yellow fever virus, which has a remarkable mortality level (being 6.8 times higher in Caucasians), could have generated protective genetic variants against itself, HCV and DENV[34]. Not only infection-related, but also metabolic genes are under intense selective pressure, and both can interact as in the case of the low-activity alleles of glucose-6-phosphatase dehydrogenase providing reduced risk to malaria infection[35]. Here we show that *OSBPL10* and *RXRA* genes, involved in lipid metabolism, are suggestive of African-conferring resistance to the development of severe dengue disease in admixed Cubans.

It is well known that Africans have a lower atherogenic lipid profile, characterized by low triglyceride (T), total cholesterol, and LDL levels, and high high-density lipoproteins (HDL) levels, when compared to Europeans[36]. Signs of selection were identified in West Africans for *APOL1* (intervening in HDL) and *CD36* (intervening in LDL) genes, probably driven by pathogen resistance: *APOL1* kills parasites through uncontrolled osmotic swelling, and *CD36* has been associated with susceptibility to malaria[37]. Our evidence adds two new genes to the differential lipid profile between Africans and Europeans, which play a role in infectious resistance. We detected signs of positive selection on the *OSBPL10* gene, while balancing selection may have generated additional polymorphisms regulating *RXRA* expression in Africans.

A direct link between lipid profile changes and dengue infection progression has been made, so that lipids may be used as predictors of clinical outcome[38]: increased T and

very-LDL were observed in severe dengue; increased HDL was observed in dengue with warning signs and severe dengue; the lowest LDL values were found in severe dengue. Probably this link results from lipids being essential to DENV entrance and replication[39]. Flavivirus RNA synthesis and replication occur on an extended network of modified endoplasmic reticulum (ER) membranes, and then maturation takes place in the ER and Golgi complex, followed by release of the mature virus from the cell[40]. These are the main cellular components where OSBPs play a key role on cholesterol and other sterols (as oxysterols) homeostasis[41]. Interestingly, replication of HCV and poliovirus has been shown to be dependent on OSBP, which also guaranteed the membranous web integrity, being recruited there in a PI4-kinase dependent manner, but DENV did not[42]. The facts that *OSBPL10* is associated with dengue in Cuba and that this protein's ligand is probably PI3P and not PI4P may be indicating a specificity binding mechanism between OSBP proteins and viral products. We also found, in Cuban dengue cohorts, African-related signs of association with kinases, a group of enzymes that deserves future research.

If the *OSBPL10* gene is involved in signaling and transport of lipids, *RXRA* plays a role in the second part of the cholesterol homeostasis through the LXR/RXR activation pathway, in hepatocytes and macrophages. The *SREBP* gene and oxysterols link these two parts[43]. When cholesterol is low, SREBP precursors are transported by SCAP proteins to the Golgi complex, where they are proteolytically processed, and then its amino-terminal domain migrates to the nucleus, binding and activating transcription of genes encoding enzymes required for the synthesis (*HMG CoA reductase*) and uptake (*LDLR*) of cholesterol. Under those conditions, LXR-RXR heterodimers recruit corepressor complexes and actively repress transcription of genes that mediate cholesterol efflux, such as *ABCA1*, and degradation of *LDLR*, leading to higher cellular concentrations of cholesterol. When cholesterol is in excess, as occurs in infection, *SREBP* precursors are sequestered in the ER by the action of cholesterol, desmosterol and oxysterols. In this situation, the consequent lowering of cellular concentrations of cholesterol is reached by oxysterol and desmosterol binding to *LXR*, which forms heterodimers with *RXRA* in order to control the transcription activation of many genes regulating cholesterol metabolism, biosynthesis, transport and efflux, lipoprotein synthesis and lipogenesis (Fig. 5).

In macrophages, *RXRA* regulates the integration of immune functions and lipid metabolism, thus defining the clinical outcome of dengue infection. Indeed, macrophages are key DENV targets cells and also the principal source of pro-inflammatory mediators linked to severe dengue disease. LXR/RXRA heterodimers and sumoylated LXR inhibit the *NF-kB* transcription factor complex, which *per se* acts upon inflammatory mediators (Fig. 5)[44]. But a negative control of LXR/RXRA dimers is exerted by *IRF3*, which is activated by viruses and bacteria that entered the cell through *TLR3* and *TLR4* receptors, deregulating the cholesterol control and allowing the action of *NF-kB* transcription factor. It has been demonstrated that during dengue infection[25] there is down-regulation of *RXRA* expression in an *IRF3*-dependent manner in the host cell, to achieve optimal *IFN* expression, and after infection *RXRA* expression resumes, suppressing the type I IFN induction. Therefore, *RXRA* not only maintains the basal type I IFN and modulates the host antiviral response, but also regulates the antiviral inflammatory response. All these cascades controlling lipid homeostasis, inflammation process and immune response are highly dynamic, with several positive and negative feedback checks.

The involvement of *RXRA* genes in infection by DENV or other viruses is supported by many other types of evidence: the a/b subdomain of the non-structural viral protein NS1 wing resembles the helicase domain of *RIG-I*, the retinoic acid-inducible gene I, which is bound by RAR-RXR (*RAR* - retinoic acid receptor) heterodimers[45]; Epstein-Barr virus BZLF-1 interacts with the ligand-binding domain of *RXR* and *RAR*, repressing the transcriptional activity of *RXR*[46]; HCV core protein positively interacts with *RXR* in its DNA-binding domain[47].

Remarkably, our results offer a comprehensive explanation for various independent observations made in association with dengue and other related viruses. *RXRA* forms heterodimers with *VDR*, which has been identified as associated with dengue protection in the Vietnamese population[48]. The RXRA-VDR heterodimers negatively control the expression of several immune function genes[49]. Also, *PLCE1*, with genetic variants detected in Vietnamese children as protective against DSS[6], interacts with *RXRA* in the PPARA/RXRA activation pathway, positively controlling the expression of genes again related with lipid metabolism[23]. Thus, in addition to the advanced hypothesis that *PLCE1* protection against DF is related to its role in maintaining the normal vascular endothelial cell barrier function[6], its involvement in the lipid metabolism should be

functionally assayed. Perhaps flavivirus resistance arose in Asians through *VDR* and *PLCE1* selection, while *RXRA* selection provided resistance in Africans, all related with central lipids and cytokines pathways. Even the immune system genes associated with dengue illness, such as *TNF-α*, *IL-10*, *TGF-β1*, *FcγRIIa* and *CD209*[4], can also be related with the *RXRA* gene in several pathways, as their expression is controlled by dimers formed by RXRA and other nuclear factors.

The genomic confirmation of the protection conferred by *OSBPL10*, *RXRA* and related lipid metabolism against dengue illness supported in this study point out potential therapeutic applications, which are extensive to a large plethora of diseases, including inflammatory, other infectious, cardiovascular and metabolic diseases. The first-generation of synthetic ligands of *LXR* were promising, but they increased dramatically hepatic lipogenesis[50]. The hypothesized specific PI3P activation of *OSBPL10* involvement in dengue would allow specific control of this flavivirus replication. This could be tested through the various PI3K (phosphorylates PI2P in PI3P) inhibitors being developed in cancer treatment[51]. In addition, the African protection against DF through kinases supports pursuing the development of kinase inhibitors as they seem to be able to block DENV assembly[21].

Methods

Cuban samples.

The Republic of Cuba has 15 provinces. Havana (La Habana) is the capital city located at the western part of the country with 2 million inhabitants, while Guantanamo city is located at the eastern part with over 200 thousand inhabitants. The Eastern and Western sides of the Cuban island show notable differences in the historic process of settlement and, consequently, in the demographic significance of the different ethnic groups[18]. The natural DENV history in Cuba is well documented, and DENV experience is also different among the Cuban provinces. After an absence of 40 years, DENV1 was reported in 1977 and transmitted to nearly one-half of the Cuban population. Four years later, DENV 2 (Asian origin) infected approximately 25% of the population, and a large DHF/DSS epidemic occurred. Sixteen years later, in 1997, another Asian DENV 2 virus entered the country, producing a located epidemic in the municipality of Santiago de Cuba. Later, in 2001, a new serotype, DENV 3 (Asian genotype), was detected in Havana city[52]. In 2006, the circulation of DENV 4 was reported in Havana, while DENV 3 affected Guantanamo[53]. Considering this, and the difference in ethnic components, we included samples from the two cities, in a total of 274: 146 from Havana city and 128 from Guantanamo city. The case groups were collected during the 2006 outbreak, classified according to WHO[54] and included: 67 subjects from Havana with confirmed dengue infection by DENV 4, 36 clinically classified as DF and 31 as DHF; and 70 subjects from Guantanamo with confirmed dengue infection by DENV 3, 41 classified as DF and 29 cases as DHF. Dengue infection was confirmed by dengue IgM detection in serum collected at day 6 of fever onset, as well as virus isolation in *Aedes albopictus* cell line and RT-PCR in samples collected in the first four days of fever[55,56].

A screening for identifying dengue asymptomatic cases was conducted during the peak of the 2006 outbreak. Samples from healthy adult individuals, without any dengue clinical symptoms, relatives or neighbours of confirmed dengue patients from three selected blocks located in neighbourhoods with a high incidence of the disease, were tested to determine asymptomatic dengue infection in the same way as the symptomatic cases. Subjects were daily visited and checked for any clinical dengue symptoms over a 15 day period. The ones that remained symptom free, and who were PCR or IgM positive, were considered as asymptomatic cases. A total of 32 asymptomatic infected

individuals from Havana and 16 from Guantanamo, that were unrelated to the individuals included in the case groups, were thus included in this work. Forty seven samples from blood donor individuals from Havana and 42 from Guantanamo were included as population controls.

Havana groups are referred as HH (dengue haemorrhagic fever), HF (dengue fever), HA (asymptomatic infection) and HC (controls). Similarly, GH, GF, GA and GC codes were used for the Guantanamo groups.

The study was conducted according to the Helsinki Declaration as a statement of ethical principles for medical research involving human subjects[57], and was approved by the Institutional Ethical Review Committee of the Institute of Tropical Medicine Pedro Kourí (IPK) and by the Ethical Committee of the Cuban National Academy of Sciences. Written informed consent was obtained from all individuals.

DNA extraction, GWAS genotyping and data quality control. Peripheral venous blood (10 ml) was collected from each individual in tubes containing acid citrate dextrose solution as anticoagulant. Genomic DNA was extracted using a commercially available kit (Qiagen, Valencia, CA) and stored at −20°C.

Genotyping was performed at the Eukaryote Genotyping Platform, Genopole, Pasteur Institute, by using the Illumina Human Omni 2.5 chip, which contains 2,379,855 SNPs. Genotype calls were obtained after scanning on the Illumina IScan Microarray System, by using the Genome Studio software. Quality control was performed in PLINK[58], and SNPs with more than 5% missing genotypes, minor allele frequency (MAF) below 1%, and Hardy-Weinberg equilibrium (HWE) deviation p-values of less than 0.001 were filtered out from downstream analyses. All samples were screened for missingness, where highest observed missingness value was 1.93%. We also checked for outliers in principal-component analysis (PCA), and for samples which showed an excess rate of heterozygous *loci* in comparison with the expected rate from the allele frequencies in the population, or had evidence of being a second-degree relative or closer to another sample in the study (identity by descent >30%; or identity by state >90%). All studied samples passed these criteria. SNPs located in X and Y chromosomes and in mitochondrial DNA were removed from the analyses, leading to a final account of 1,922,396 autosomal SNPs.

Global population structure. In order to ascertain the global ancestry background, the Cuban samples were compared with populations from Europe, Africa, Latin America and East Asia, from the 1000 Genomes project[59] and another worldwide dataset[60] (Supplementary Table 1), so that the total final dataset contained 389,574 common SNPs.

The genomic proportions of European, African, Native American and Asian ancestry were evaluated in all Cuban groups using the program ADMIXTURE[61], which provides a maximum likelihood estimation of the population structure. For this analysis, SNPs were pruned for pairwise linkage disequilibrium, removing SNPs with $r^2 > 0.4$ with another SNP within a 50-SNP window, ending up with a total of 77,782 SNPs. We ran ADMIXTURE for several numbers (from 2 to 10) of ancestral populations, *K*. The optimal *K* was estimated through cross-validation of the logistic regression.

PCA infers worldwide axes of human genetic variation from the allele frequencies of various populations. It was carried out by using the *smartpca* tool, and the statistical significance of each PC was evaluated through the Tracy-Widom statistics available at the *twstats*, both included in the EIGENSOFT package[62].

Global ancestry influence in dengue infection outcome in Cuba. Two-tailed Wilcoxon rank-sum test (non-parametric test, not requiring normal distribution) was applied to assess the significance between ancestry proportions in the Cuban groups. Results were displayed in box plots, by using Origin 7 software (www.originlab.com). We used R to apply an iterative model, by considering the following variables: African and Native American ancestries as discontinuous variables; Havana and Guantanamo locations as binary variables; and asymptomatic and DHF phenotypes as binary variables.

Fine-matched population structure correction and association analysis. Based on the information from the global admixture, we performed a fine-matched correction for population structure, avoiding spurious associations resulting from differences in the ancestral background. Comparison groups were organized by matching the paired individuals by the global African ancestry, so that the pair did not differ by more than 2% in this ancestry. Havana and Guantanamo individuals were included together, but as far as possible the comparison pair was from the same city, and asymptomatic infected individuals were considered first than controls as the non-dengue patients in the

comparison pair. We obtained 54 asymptomatic/control versus DHF pairs (haemorrhagic comparison group - HCG) and 74 asymptomatic/control versus DF pairs (fever comparison group - FCG). A 111 pair group mixing disease phenotypes was also tested for asymptomatic/control versus DHF/DF (overall comparison group – OCG).
Single-locus allelic association analyses were carried out in PLINK[58] through the χ^2 statistics. P-values were displayed as $-\log_{10}$ in Manhattan plots, obtained in HaploView.

Admixture mapping. Local ancestry analysis was performed along the mosaic admixed chromosomes, by using the algorithm RFMix[63]. We used two parental populations from the 1000 Genomes database to represent the European and African components, 50 Italian samples and 50 Yoruban samples respectively, and phased them together with the Cuban samples using SHAPEIT[64] with default options and the fine scale genetic map from HapMap phase II. We performed a test in chromosome 22 with the Spanish sample from 1000 Genomes database, instead of the Italian, and confirmed that results were identical.
The proportions of the two ancestries were calculated for every position along each chromosome in the admixed Cuban groups. In order to identify regions which have a significantly high proportion of the African parental ancestry, we estimated the difference in African proportion in each comparison group, and followed others[65] in considering as significant the genomic regions in which that ancestry was outside the range defined by the genome mean ± 3SD.

Gene expression analysis in Cuban patients. A group of 20 patients were recruited from those admitted at the Salvador Allende Hospital, in Havana in 2014, with suspected dengue acute illness (clinical diagnosis was made by an experienced physician in dengue illness). Dengue infection was confirmed as described before. Patients were examined daily during hospitalization by physicians and nurses experienced in dengue management. General signs and symptoms as well as warning signs were recorded daily from recruitment to discharge. Three serial peripheral mononuclear cell (PBMC) samples were kinetically collected during hospitalization, at days 3, 7 and 30, after symptoms onset. Whole blood (10 mL) was drawn from the median cubital vein of study subjects into EDTA-containing tubes. PBMCs were obtained by Histopaque-1077 (Sigma-Aldrich, UK) density gradient centrifugation and frozen in RNA protect cell reagent (Qiagen, Hilden, Germany).

DNase-treated total RNA was isolated from PBMC using the RNeasy Mini kit (Qiagen, Hilden, Germany) and evaluated by using the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA). Expression levels of *OSBPL10* and *RXRA* were determined by RT-PCR at LightCycler Carousel-Based System (LightCycler 2.0 instrument, Roche), using LightCycler RNA Master SYBR Green I kit (Roche) for One-Step RT-PCR, using the specific forward and reverse primers for the studied genes and the β-actin housekeeping gene (Supplementary Table 2).

Each sample was subjected to RT-PCR in duplicate and the mean values of the duplicates were used for subsequent analysis. Ct values of studied genes were normalized to an average Ct value of the corresponding housekeeping gene, and the relative expression of each representative was calculated. The assay specificity was evaluated by melting curve analysis.

Statistical analysis was performed using the non-parametric Wilcoxon-Mann-Whitney U mean rank test for quantitative variables. Data were displayed as mean and standard deviation. A 5% level of statistical significance was considered.

Confirmation of mRNA expression, potential functional role and positive selection in public databases. We checked the expression of the significant genes detected in our study in several public datasets. One dataset consisted in 465 RNASeq performed in the lymphoblastoid cell lines from the 1000 Genomes project[28], from European (descendants living in Utah, Finns, British and Italians) and African Yoruba (Nigerian) individuals. The quantification expressed in reads per kilobase of exon per million reads mapped (RPKM) was extracted from ArrayExpress (E-GEUV-1). We also used the data for the whole blood transcriptome in a Thai dengue dataset[20], composed of the following samples: nine healthy controls; 28 samples collected between days 2 and 9 after onset of symptoms (acute illness) from secondarily infected patients, divided in 18 DF and 10 DHF; 19 samples (13 DF and 6 DHF) collected at convalescence, four weeks or later after discharge. The expression data from this publication were obtained from the GEO profiles (GDS5093). We used a linear discriminant analysis effect size (LEfSe) method[66] for high-dimensional class comparisons in the African-related and in the non-African-related protective genes detected in the association test, and also in the African enriched genes identified in RFMix test, in the three comparison groups. Additionally, GSEA software[32] was used to run a gene set enrichment analysis for the LXR/RXR activation pathway in macrophages (information compiled from literature

and ingenuity database; Supplementary Table 20) for the Thai dengue dataset[20]. The pathway was divided in three sets of genes: lipid metabolism, LXR/RXR activation and NF-kB activation.

We checked if significant SNPs or haplotypes could be located on promoter regions by using the tool EPDNew human version 003[29]. For checking the location of several regulatory regions, including promoters, enhancers and motif binding regions, we used HaploReg version 2[67], which includes an expanded library of SNPs, motif instances, enhancer annotations, eQTLs and frequency information based on the 1000 Genomes Phase 1 individuals.

We also used Haplotter database[31] to explore the evidence for recent positive selection for the candidate genes. This database reports information on diversity across the globe and positive selection through iHS measures, which is has good power to detect selective sweeps at moderate frequency (50–80%). By using the selscan package[68], we applied another measure of positive selection to all genome data in the HCG, XP-EHH[69], that is most powerful for selective sweeps above 80% frequency.

References

1 Guzman, M. G. & Harris, E. Dengue. *Lancet* **385**, 453-465 (2015).
2 Ross, T. M. Dengue virus. *Clin. Lab. Med.* **30**, 149-160 (2010).
3 Grange, L. *et al.* Epidemiological risk factors associated with high global frequency of inapparent dengue virus infections. *Front. Immunol.* **5**, 280 (2014).
4 Coffey, L. L. *et al.* Human genetic determinants of dengue virus susceptibility. *Microbes Infect.* **11**, 143-156 (2009).
5 Garcia, G. *et al.* Association of MICA and MICB alleles with symptomatic dengue infection. *Hum. Immunol.* **72**, 904-907 (2011).
6 Khor, C. C. *et al.* Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1. *Nat. Genet.* **43**, 1139-1141 (2011).
7 de la, C. S. B., Kouri, G. & Guzman, M. G. Race: a risk factor for dengue hemorrhagic fever. *Arch. Virol.* **152**, 533-542 (2007).
8 Agramonte, A. Notas clinicas sobre una epidemian reciente de dengue. *Rev. Med. Cirug. Cub.* **January**, 222-226 (1906).
9 Guzman, M. G. *et al.* [Dengue in Cuba: history of an epidemic]. *Rev. Cubana Med. Trop.* **40**, 29-49 (1988).
10 Halstead, S. B. *et al.* Haiti: absence of dengue hemorrhagic fever despite hyperendemic dengue virus transmission. *Am. J. Trop. Med. Hyg.* **65**, 180-183 (2001).
11 Jaenisch, T. *et al.* Dengue expansion in Africa-not recognized or not happening? *Emerg. Infect. Dis.* **20** (2014).
12 Chacon-Duque, J. C. *et al.* African genetic ancestry is associated with a protective effect on Dengue severity in colombian populations. *Infect. Genet. Evol.* **27**, 89-95 (2014).
13 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904-909 (2006).
14 Winkler, C. A., Nelson, G. W. & Smith, M. W. Admixture mapping comes of age. *Annu. Rev. Genomics Hum. Genet.* **11**, 65-89 (2010).
15 Patterson, N. *et al.* Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74**, 979-1000 (2004).
16 Galanter, J. M. *et al.* Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: the Genes-environments & Admixture in Latino Americans study. *J. Allergy Clin. Immunol.* **134**, 295-305 (2014).
17 Jeff, J. M. *et al.* Admixture mapping and subsequent fine-mapping suggests a biologically relevant and novel association on chromosome 11 for type 2 diabetes in African Americans. *PLoS One* **9**, e86931 (2014).
18 Guerra, R. *Manual de Historia de Cuba. Desde su descubrimiento hasta 1868.* (Editorial Nacional de Cuba, 1964).
19 Marcheco-Teruel, B. *et al.* Cuba: exploring the history of admixture and the genetic basis of pigmentation using autosomal and uniparental markers. *PLoS Genet.* **10**, e1004488 (2014).
20 Kwissa, M. *et al.* Dengue virus infection induces expansion of a CD14(+)CD16(+) monocyte population that stimulates plasmablast differentiation. *Cell Host Microbe* **16**, 115-127 (2014).
21 Anwar, A. *et al.* The kinase inhibitor SFV785 dislocates dengue virus envelope protein from the replication complex and blocks virus assembly. *PLoS One* **6**, e23246 (2011).
22 Fairn, G. D. & McMaster, C. R. The roles of the human lipid-binding proteins ORP9S and ORP10S in vesicular transport. *Biochem. Cell. Biol.* **83**, 631-636 (2005).
23 Motojima, K., Passilly, P., Peters, J. M., Gonzalez, F. J. & Latruffe, N. Expression of putative fatty acid transporter genes are regulated by peroxisome proliferator-

activated receptor alpha and gamma activators in a tissue- and inducer-specific manner. *J. Biol. Chem.* **273**, 16710-16714 (1998).

24 Gao, X. *et al.* A genome-wide association study of central corneal thickness in Latinos. *Invest. Ophthalmol Vis. Sci.* **54**, 2435-2443 (2013).

25 Ma, F. *et al.* Retinoid X receptor alpha attenuates host antiviral response by suppressing type I interferon. *Nat. Commun.* **5**, 5494 (2014).

26 Caromile, L. A., Oganesian, A., Coats, S. A., Seifert, R. A. & Bowen-Pope, D. F. The neurosecretory vesicle protein phogrin functions as a phosphatidylinositol phosphatase to regulate insulin secretion. *J. Biol. Chem.* **285**, 10487-10496 (2010).

27 Saeed, M. *et al.* SEC14L2 enables pan-genotype HCV replication in cell culture. *Nature* **524**, 471-475 (2015).

28 Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-511 (2013).

29 Dreos, R., Ambrosini, G., Perier, R. C. & Bucher, P. The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. *Nucleic Acids Res.* **43**, D92-96 (2015).

30 Heltemes-Harris, L. M., Willette, M. J., Vang, K. B. & Farrar, M. A. The role of STAT5 in the development, function, and transformation of B and T lymphocytes. *Ann. N. Y. Acad. Sci.* **1217**, 18-31 (2011).

31 Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).

32 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545-15550 (2005).

33 Soares, P. *et al.* The Expansion of mtDNA Haplogroup L3 within and out of Africa. *Mol. Biol. Evol.* **29**, 915-927 (2012).

34 Blake, L. E. & Garcia-Blanco, M. A. Human genetic variation and yellow fever mortality during 19th century U.S. epidemics. *mBio* **5**, e01253-01214 (2014).

35 Tishkoff, S. A. *et al.* Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293**, 455-462 (2001).

36 Goedecke, J. H. *et al.* Ethnic differences in serum lipoproteins and their determinants in South African women. *Metabolism* **59**, 1341-1350 (2010).

37 Ferrer-Admetlla, A., Liang, M., Korneliussen, T. & Nielsen, R. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol. Biol. Evol.* **31**, 1275-1291 (2014).

38 Duran, A. *et al.* Association of lipid profile alterations with severe forms of dengue in humans. *Arch. Virol.* **160**, 1687-1692 (2015).

39 Soto-Acosta, R. *et al.* The increase in cholesterol levels at early stages after dengue virus infection correlates with an augment in LDL particle uptake and HMG-CoA reductase activity. *Virology* **442**, 132-147 (2013).

40 Apte-Sengupta, S., Sirohi, D. & Kuhn, R. J. Coupling of replication and assembly in flaviviruses. *Curr. Opin. Virol.* **9**, 134-142 (2014).

41 Du, X., Brown, A. J. & Yang, H. Novel mechanisms of intracellular cholesterol transport: oxysterol-binding proteins and membrane contact sites. *Curr. Opin. Cell Biol.* **35**, 37-42 (2015).

42 Wang, H. *et al.* Oxysterol-binding protein is a phosphatidylinositol 4-kinase effector required for HCV replication membrane integrity and cholesterol trafficking. *Gastroenterology* **146**, 1373-1385.e1371-1311 (2014).

43 Spann, N. J. & Glass, C. K. Sterols and oxysterols in immune cell function. *Nat. Immunol.* **14**, 893-900 (2013).

44 Joseph, S. B., Castrillo, A., Laffitte, B. A., Mangelsdorf, D. J. & Tontonoz, P. Reciprocal regulation of inflammation and lipid metabolism by liver X receptors. *Nat. Med.* **9**, 213-219 (2003).

45 Jiang, S. Y. *et al.* Identification and characterization of the retinoic acid response elements in the human RIG1 gene promoter. *Biochem. Biophys. Res. Commun.* **331**, 630-639 (2005).

46 Sista, N. D., Barry, C., Sampson, K. & Pagano, J. Physical and functional interaction of the Epstein-Barr virus BZLF1 transactivator with the retinoic acid receptors RAR alpha and RXR alpha. *Nucleic Acids Res.* **23**, 1729-1736 (1995).

47 Tsutsumi, T. *et al.* Interaction of hepatitis C virus core protein with retinoid X receptor alpha modulates its transcriptional activity. *Hepatology* **35**, 937-946 (2002).

48 Loke, H. *et al.* Susceptibility to dengue hemorrhagic fever in vietnam: evidence of an association with variation in the vitamin d receptor and Fc gamma receptor IIa genes. *Am. J. Trop. Med. Hyg.* **67**, 102-106 (2002).

49 D'Ambrosio, D. *et al.* Inhibition of IL-12 production by 1,25-dihydroxyvitamin D3. Involvement of NF-kappaB downregulation in transcriptional repression of the p40 gene. *J. Clin. Invest.* **101**, 252-262 (1998).

50 Zelcer, N. & Tontonoz, P. Liver X receptors as integrators of metabolic and inflammatory signaling. *J. Clin. Invest.* **116**, 607-614 (2006).

51 Anderson, J. L. *et al.* Evaluation of In Vitro Activity of the Class I PI3K Inhibitor Buparlisib (BKM120) in Pediatric Bone and Soft Tissue Sarcomas. *PLoS One* **10**, e0133610 (2015).

52 Guzman, M. G. & Kouri, G. Dengue in Cuba: research strategy to support dengue control. *Lancet* **374**, 1660-1661 (2009).

53 Libel, M. Brote de dengue en Cuba, 2006. *Enfermedades infecciosas emergentes y reemergentes, Región de las Américas, Habana.* Cuba: OPS (2006).

54 WHO. *Dengue haemorrhagic fever: diagnosis, treatment, prevention and control.* 2nd edn (World Health Organization, 1997).

55 Vazquez, S., Bravo, J. R., Perez, A. B. & Guzman, M. G. [Inhibition ELISA. Its utility for classifying a case of dengue]. *Rev. Cubana Med. Trop.* **49**, 108-112 (1997).

56 Rodriguez Roche, R., Alvarez, M., Guzman, M. G., Morier, L. & Kouri, G. Comparison of rapid centrifugation assay with conventional tissue culture method for isolation of dengue 2 virus in C6/36-HT cells. *J. Clin. Microbiol.* **38**, 3508-3510 (2000).

57 Association, W. M. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA* **310**, 2191-2194 (2013).

58 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575 (2007).

59 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

60 Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104 (2008).

61 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655-1664 (2009).

62 Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

63 Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278-288 (2013).

64 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179-181 (2012).

65 Bryc, K. *et al.* Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc. Natl. Acad. Sci. USA* **107**, 786-791 (2010).
66 Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
67 Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930-934 (2012).
68 Szpiech, Z. A. & Hernandez, R. D. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* **31**, 2824-2827 (2014).
69 Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913-918 (2007).

Acknowledgements

We dedicate this paper to the memory of Professor Gustavo Kourí Flores. We thank all the Cuban individuals from Havana and Guantanamo cities who agreed to participate in this study. The authors express thanks to Dr Torreblanca from the Centre of Hygiene and Epidemiology of Guantanamo City, to Dr. Raiza Martinez from the Salvador Allende Hospital, and to Dr. Rosa Martinez and technicians Barbara Marrero and Jose Rodriguez from the Pedro Kourí Institute, for their support in the samples collection. The authors also thank the technician Laure Lemée, from the Eukaryote Genotyping Platform Institut Pasteur, for her helpful support with the genotyping. Thanks also to Dr. Daniel Limonta for his helpful reviewing of this paper. The research leading to these results has received funding from the European Commission Seventh Framework Programme [FP7/2007-2013] for the DENFREE project under Grant Agreement no. 282378. PT has a PhD grant from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 290344 (EUROTAST) and MS a PhD grant from FCT (The Portuguese Foundation for Science and Technology - SFRH/BD/95626/2013). IPATIMUP integrates the i3S Research Unit, which is partially supported by FCT, FEDER and COMPETE (PEst-C/SAU/LA0003/2013).

Author contributions

BS conceived the study, collected the Cuban samples and leaded the performance of all lab analyses with the collaboration of GG, ABP, EA, MA and DR. BR performed GWAS genotyping and data quality control. PT performed the computational analysis

under supervision of PS and LP. MO and BC conducted the GSEA analysis of the transcriptome dataset under supervision of PS and LP. DCS provided statistical expertise. AS, LP and MGG designed the project. BS, PT and LP wrote the manuscript, which was revised and approved by all authors.

Table 1 | Odds ratios of the African ancestry influence in DHF phenotype when compared to asymptomatic subjects, in Cuba in general, only Havana city and in Colombia.

	Odds ratio		
	1% African ancestry increase	50% African ancestry increase	100% African ancestry increase
Cuba	0.979	0.396	0.151
Havana	0.920	0.045	0.012
Colombia*	0.962	0.204	0.042

* From (Chacon-Duque et al. 2014)

Table 2 | Odds ratios of the African *OSBPL10* haplotype and *RXRA* alleles in DHF when compared with asymptomatic/control and tests (identified by YES) where statistical significant evidence was detected for each gene. HCG means haemorrhagic comparison group (54 asymptomatic/control versus DHF pairs).

Gene	Haplotype/SNP	Position	Haplotype/Allele	Odds ratio	OR 95% confidence interval	Association (HCG)	Admixture mapping (HCG)	Positive selection (XP-EHH and iHS)	Expression changes in Cuban patients RT-PCR	Expression changes in Thai transcriptome dataset	Expression changes between African and European haplotypes/genotypes in 1000 Genomes transcriptome
OSBPL10	African haplotype			0.2486	[0.1329-0.4650]	YES		YES	YES	YES	YES
	rs4600849	32027672									
	rs11129475	32030544									
	rs6419811	32031135									
	rs11718700	32033248									
	rs975406	32035587									
	rs7639637	32036042	CTGCCC								
RXRA	rs12339163	137205188	G	0.3577	[0.1666-0.7677]		YES		YES	YES	
	rs62576287	137214888	C	0.1028	[0.0128-0.8262]						
	rs3118593	137426334	A	0.4356	[0.2544-0.7707]						
	rs4262378	137515156	G	0.4133	[0.2378-0.7183]						
	rs4424343	137515158	A	0.4299	[0.2441-0.7569]						

Figure legends

Figure 1 | The global ancestry in Cuba and its influence on the susceptibility to dengue. (a) ADMIXTURE results for K=4. HC: Havana controls, HA: Havana individuals with asymptomatic infection, HH: Havana DHF cases, HF: Havana DF cases, GC: Guantanamo controls, GA: Guantanamo individuals with asymptomatic infection, GH: Guantanamo DHF cases, GF: Guantanamo DF cases. **(b)** Box plots for the African ancestry in the Cuban groups: controls; individuals with asymptomatic infection; DF (dengue fever); DHF (dengue haemorrhagic fever). The boxes represent the interquartile range and the whiskers are the 5% and 95% quartiles. The significant p-values for the two-tailed Wilcoxon rank-sum test between pairs of groups are displayed; non-significant ones are not displayed. **(c)** The DHF predicted probability curves in function of the African ancestry in Havana (blue) and Guantanamo (pink), by comparison with asymptomatic.

Figure 2 | The relevant region on chromosome 3 containing the *OSBPL10* gene. (a) Manhattan plot for the association analysis in the 54 fine-matched population structure corrected Cuban pairs of asymptomatic/control versus DHF subjects. **(b)** The region on chromosome 3, with the haplotype defined by the six significantly associated SNPs indicated by the red box. Genes on the forward sense are indicated in blue; genes on the reverse sense are indicated in light brown. **(c)** Worldwide frequency of the African (blue), European (red) and other (grey) *OSBPL10* haplotypes for populations of the 1000 Genomes project, and also for asymptomatic/control and DHF in Cuba. **(d)** mRNA expression for homozygous genotypes for African and European *OSBPL10* haplotypes in the 1000 Genomes project transcriptome information.

Figure 3 | The *RXRA-COL5A1* region with the most significant SNPs highlighted. Obtained with LocusZoom tool, by using recombination rate information from Yoruba. The symbols above the rule represent significant SNPs, for the Cuban data (the pink triangles), the African comparison between individuals having low (n=6; lower than 10 RPKM) and high (n=8; higher than 20 RPKM) *RXRA* expression (green circles), the same for European individuals (n=42 and n=39, respectively; blue squares).

Figure 4 | Gene expression for *RXRA* and *OSBPL10* in Cuban dengue patients along the course of disease. Data is shown for all Cuban patients, Cuban patients with warning signs, and Thai transcriptome dataset for whole genome [20].

Figure 5 | The LXR/RXR activation pathway in macrophages. Englobing the lipid metabolism, LXR/RXR activation and NF-kB activation. Information was collected from the Ingenuity database (https://targetexplorer.ingenuity.com/index.htm) and publications cited in the Discussion section [43,50]. Red lines with block in the end mean inhibition; arrows mean activation. The precise mechanism by which *OSBPL10* is involved in the transport of lipids between membranous organelles and as signal detector of cholesterol or oxysterols is still under investigation (see Discussion section).

Figure 6 | GSEA analysis in DF vs convalescent subjects, and DHF vs convalescent subjects in the Thai transcriptome dataset [20]. ES stands for enrichment score, reflecting the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes. NES is the normalized enrichment score, accounting for differences in gene set size, in correlations between gene sets and the expression dataset. FDR is the false discovery rate, the estimated probability that a gene set with a given NES represents a false positive finding (FDR<25%, meaning that the result is valid 3 out of 4 times, are highlighted in bold). The genes identified as up-regulated in each test are marked by the green shadow.

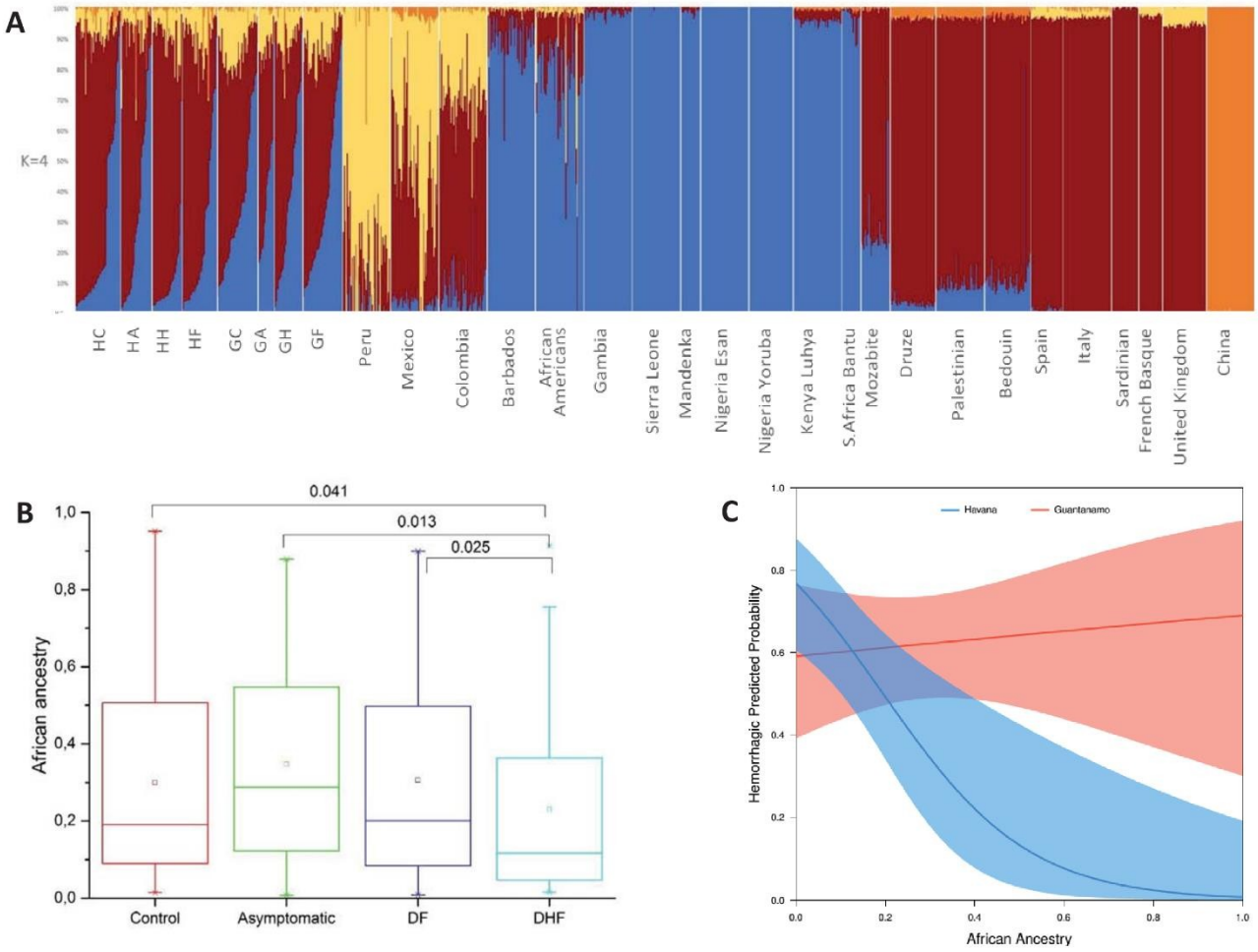

Figure 1

Figure 1

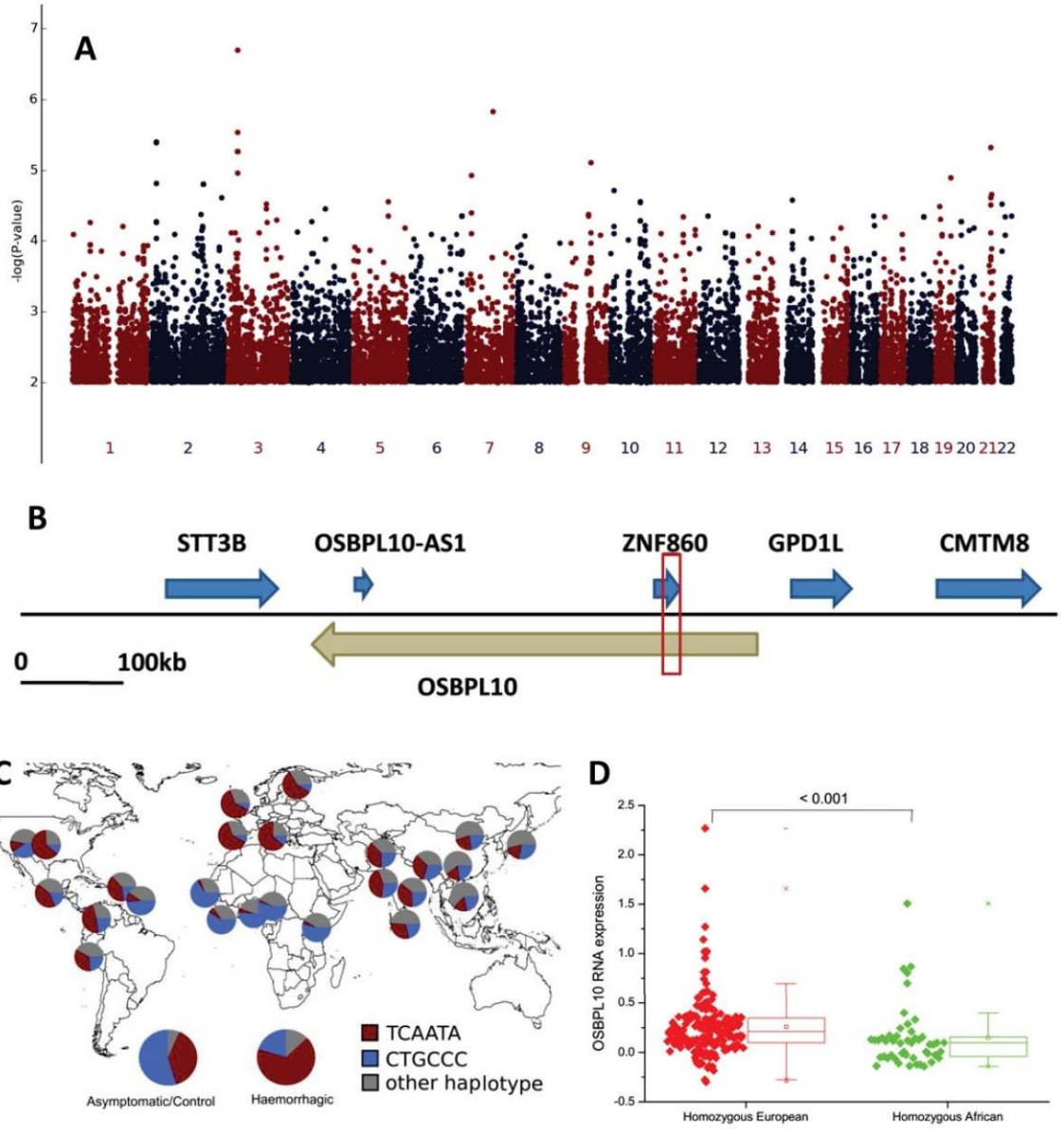

Figure 2

Figure 2

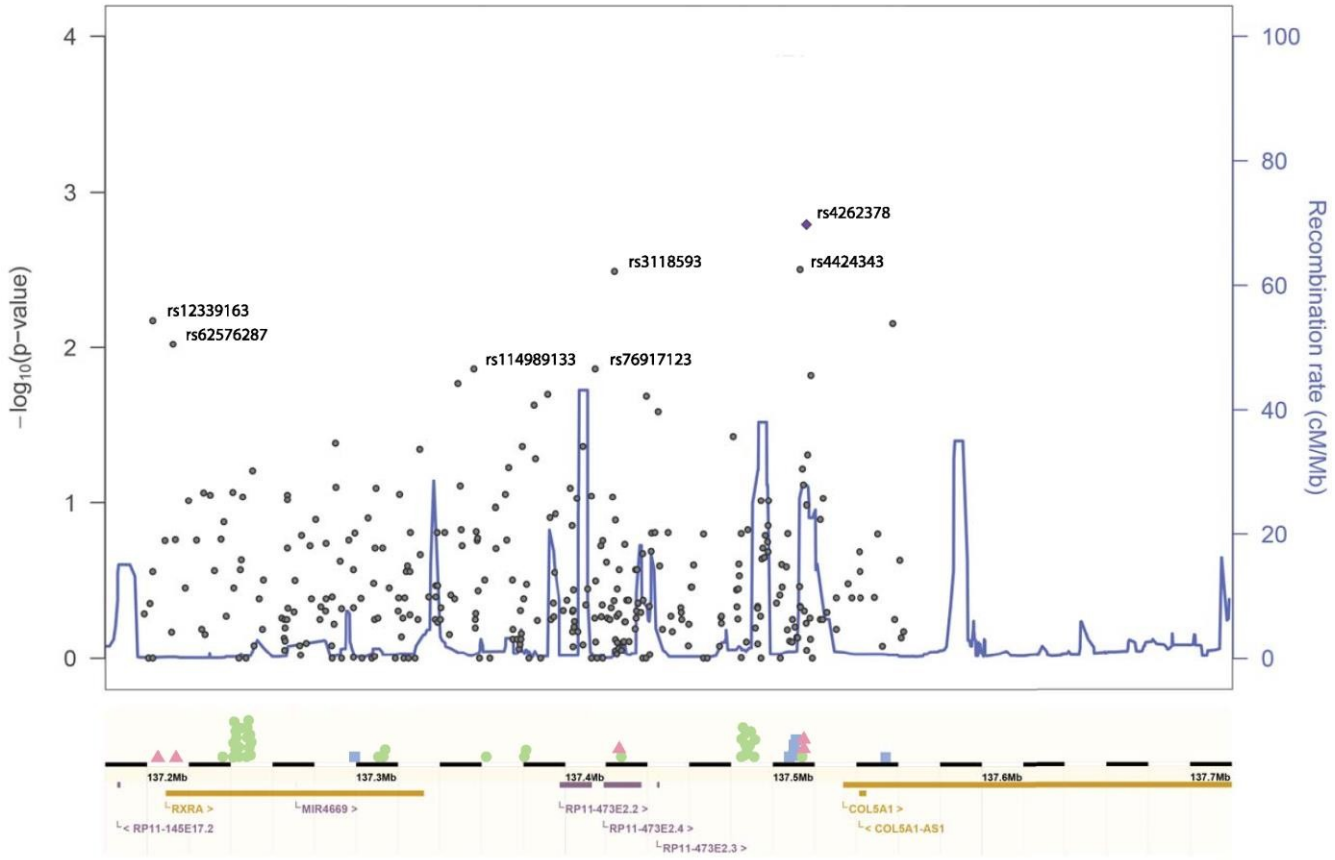

Figure 3

Figure 3

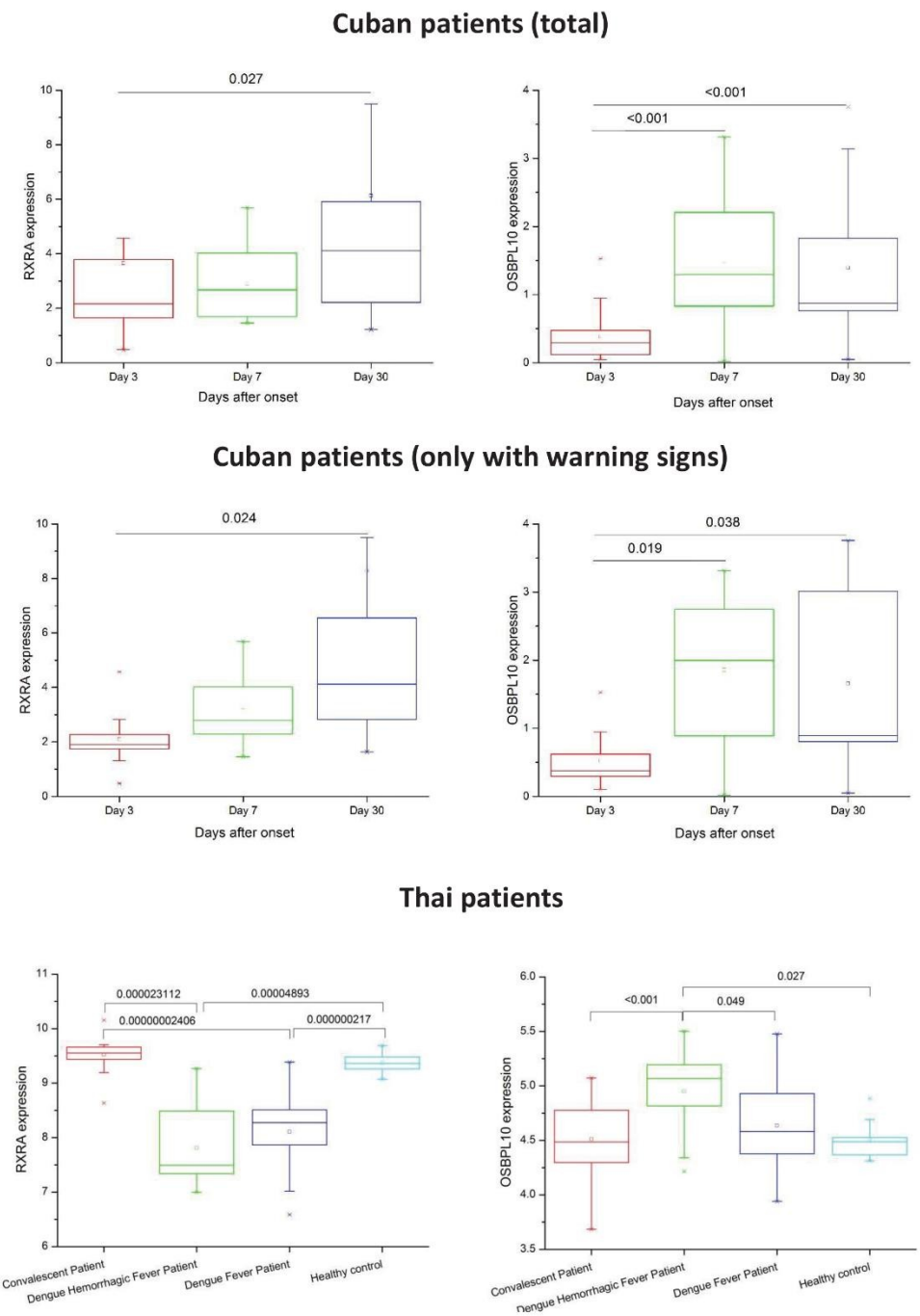

Figure 4

Figure 4

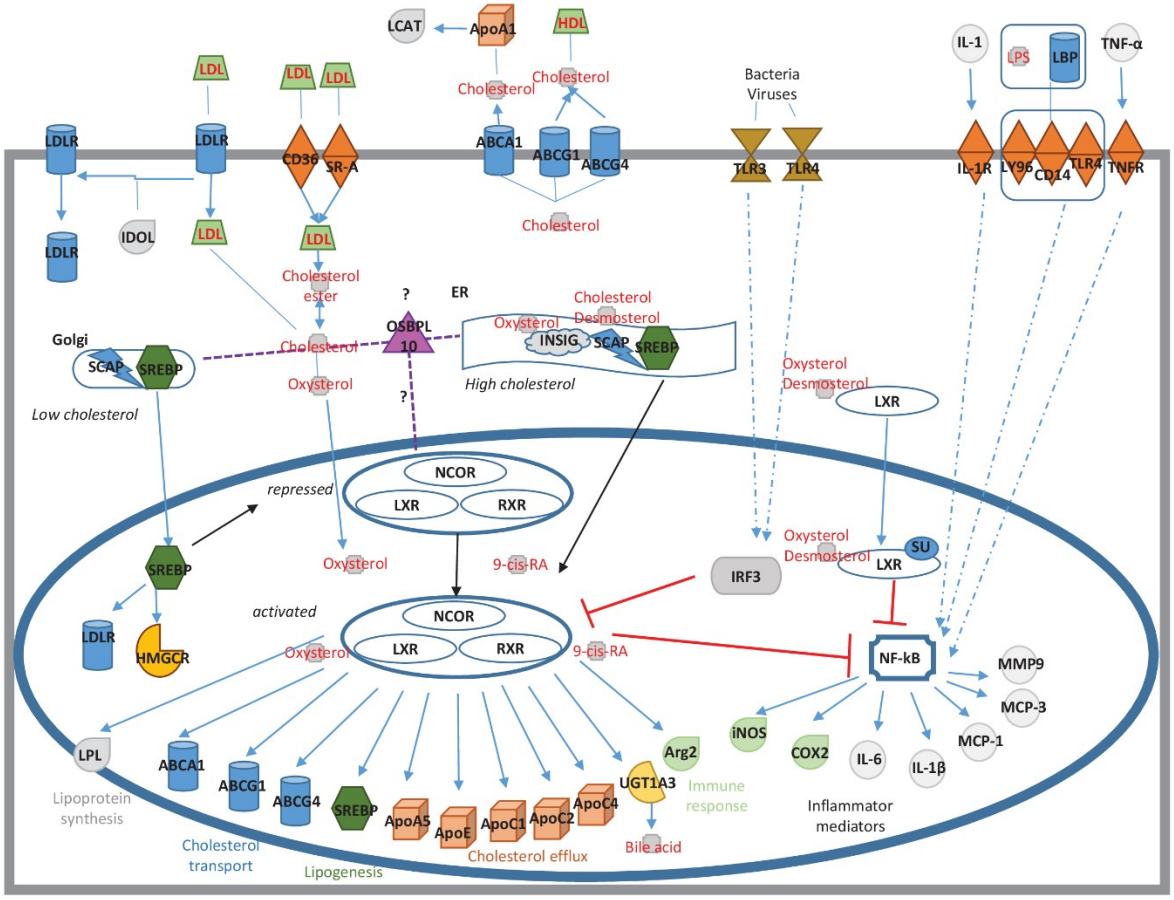

Figure 5

Figure 5

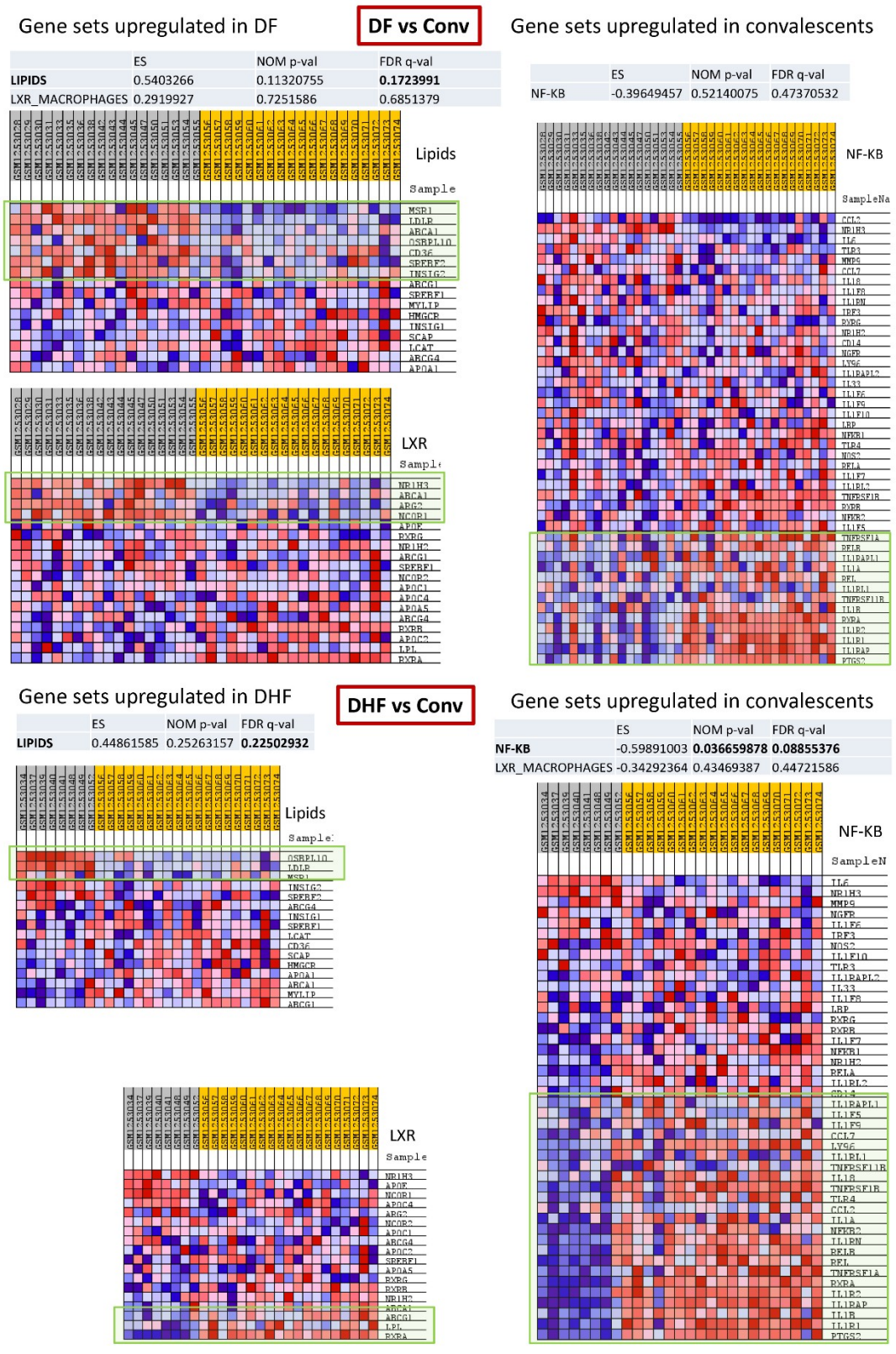

Figure 6

Figure 9 - PPAR signalling pathway. *Adaped from KEGG: Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/).*Figure 6

3.3 Paper III

Genetic Stratigraphy of Key Demographic Events in Arabia.

PloS One, *10*(3), e0118625.

PLOS | ONE

RESEARCH ARTICLE

Genetic Stratigraphy of Key Demographic Events in Arabia

Verónica Fernandes[1,2,3], Petr Triska[1,2,4], Joana B. Pereira[1,2,3], Farida Alshamali[5], Teresa Rito[2], Alison Machado[2], Zuzana Fajkošová[2,6], Bruno Cavadas[1,2], Viktor Černý[6], Pedro Soares[2], Martin B. Richards[3,7☯], Luísa Pereira[1,2,8☯*]

1 Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal, **2** Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto, Portugal, **3** School of Biology, Faculty of Biological Sciences, University of Leeds, Leeds, United Kingdom, **4** Instituto de Ciências Biomédicas da Universidade do Porto (ICBAS), Porto, Portugal, **5** General Department of Forensic Sciences and Criminology, Dubai Police General Headquarters, Dubai, United Arab Emirates, **6** Archaeogenetics Laboratory, Institute of Archaeology of the Academy of Sciences of the Czech Republic, Prague, Czech Republic, **7** Department of Biological Sciences, School of Applied Sciences, University of Huddersfield, Huddersfield, United Kingdom, **8** Faculdade de Medicina da Universidade do Porto, Porto, Portugal

☯ These authors contributed equally to this work.
* lpereira@ipatimup.pt

OPEN ACCESS

Citation: Fernandes V, Triska P, Pereira JB, Alshamali F, Rito T, Machado A, et al. (2015) Genetic Stratigraphy of Key Demographic Events in Arabia. PLoS ONE 10(3): e0118625. doi:10.1371/journal.pone.0118625

Academic Editor: Gyaneshwer Chaubey, Estonian Biocentre, ESTONIA

Received: August 1, 2014

Accepted: January 21, 2015

Published: March 4, 2015

Data Availability Statement: All whole-mtDNA sequences generated in this work are available in GenBank database, accession numbers KP316996-KP317078.

Funding: FCT, the Portuguese Foundation for Science and Technology, supported this work through the research project PTDC/CS-ANT/113832/2009 and the personal grants to V.F. (SFRH/BD/61342/2009), J.B.P. (SFRH/BD/45657/2008), and P.S. (SFRH/BPD/64233/2009). P.T. has a PhD grant from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no.

Abstract

At the crossroads between Africa and Eurasia, Arabia is necessarily a melting pot, its peoples enriched by successive gene flow over the generations. Estimating the timing and impact of these multiple migrations are important steps in reconstructing the key demographic events in the human history. However, current methods based on genome-wide information identify admixture events inefficiently, tending to estimate only the more recent ages, as here in the case of admixture events across the Red Sea (~ 8–37 generations for African input into Arabia, and 30–90 generations for "back-to-Africa" migrations). An mtDNA-based founder analysis, corroborated by detailed analysis of the whole-mtDNA genome, affords an alternative means by which to identify, date and quantify multiple migration events at greater time depths, across the full range of modern human history, albeit for the maternal line of descent only. In Arabia, this approach enables us to infer several major pulses of dispersal between the Near East and Arabia, most likely via the Gulf corridor. Although some relict lineages survive in Arabia from the time of the out-of-Africa dispersal, 60 ka, the major episodes in the peopling of the Peninsula took place from north to south in the Late Glacial and, to a lesser extent, the immediate post-glacial/Neolithic. Exchanges across the Red Sea were mainly due to the Arab slave trade and maritime dominance (from ~ 2.5 ka to very recent times), but had already begun by the early Holocene, fuelled by the establishment of maritime networks since ~ 8 ka. The main "back-to-Africa" migrations, again undetected by genome-wide dating analyses, occurred in the Late Glacial period for introductions into eastern Africa, whilst the Neolithic was more significant for migrations towards North Africa.

317184. The authors also thank the Leverhulme Trust (research project grant 10 105/D) (to M.B.R.) and the DeLaszlo Foundation (to M.B.R./P.S.) for support. The Instituto de Patologia e Imunologia Molecular da Universidade do Porto is an Associate Laboratory of the Portuguese Ministry of Science, Technology, and Higher Education and is partially supported by FCT. VČ was supported by the Grant Agency of the Czech Republic (Grant no. 13–37998S-P505). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

The issue of admixture in human populations is normally addressed by genome-wide (GW) studies, and several approaches have been developed to date admixture events [1,2,3,4,5]. Admixed populations bear chromosomes with segments of DNA from all contributing source groups, the size of which decreases over successive generations until recombination renders them undetectably short. Several algorithms attempt to date admixture events by inferring the size of the nuclear ancestry segments, and these can work well when dating recent episodes in human history, such as the sub-Saharan African input into the New World [6], but they fail to detect several known episodes that took place at earlier times, such as the African input into Iberia [1] and genetic exchanges across the Red Sea [7]. Simulations with the suite of methods available at the ADMIXTOOLS package indicated that these methods could detect admixture events as early as 500 generation ago, but real data did not allow the tracing of such old events [8]. A recent improved algorithm, called GLOBETROTTER, has been used to tackle the detection of the co-occurrence of several mixture events by decomposing each chromosome into a series of haplotypic chunks and then analysing each chunk independently [3], but the problem of detecting ancient events remains. Its application to the systematic screening of worldwide admixture events was able to reveal around 100 events, but all occurring over only the past 4,000 years [3].

The uniparental markers, characterised by the absence of recombination, do make possible the inference of ancestry for the mitochondrial genome and non-recombining, male-specific portion of the Y chromosome (mtDNA and MSY, respectively), and the dating of some demographic events (those which leave a signature in the genealogy), provided that a mutation rate of these molecules is reliably established. For the mtDNA, in the last couple of years, the application of various methods has led to quite reliable mutation rates with which to convert genetic diversities into time [9,10], while the MSY remains prone to more uncertainty [11], although promising advances are being achieved with whole Y chromosomal mutation rate calibrations [12,13,14].

At the same time, it is important to emphasize that the age of an mtDNA haplogroup cannot be directly associated with a migration event, as the diversity that has arisen in the source population, predating the migration event, would be included in the measurement. Founder analysis is an attempt to overcome this limitation. This approach picks out founder sequence types in potential source populations and dates lineage clusters deriving from them in the settlement zone of interest. In a way, the founder analysis allows us to reconstruct the stratigraphy of the migration events responsible for making up a population genetic pool, analogous to the archaeological reconstruction of the history of a site by the analysis of its sequential layers [15,16,17,18].

Some authors have been critical of dating migration events solely based upon the mtDNA evidence, arguing that maternal lineages do not necessarily represent the entire population, and are especially sensitive to drift [19]. Nevertheless, mtDNA-based conclusions for many migrations across various regions of the globe have been subsequently supported by genome-wide results [20,21], despite the limitations of the latter in dating events. In fact, the genealogical approach taken for mtDNA may overcome the effects of drift more effectively than the use of genome-wide SNPs, as we recently demonstrated in the highly-drifted Ashkenazi population: the fine characterisation of mtDNA sequences provided a detailed reconstruction of the maternal Ashkenazi pool, indicating that at least 80% of the lineages had a deep European ancestry [22], an influence not so readily identified in worldwide PCAs based on genome-wide data [23]. Thus, we suggest that for high time-depths, the mtDNA remains at present the most

informative genetic system with which to infer past migrations and estimate their time frames, allowing us to disentangle the palimpsest that results from the impact of successive migrations.

Several distinct disciplines, including climatology, archaeology and genetics, are beginning to suggest that Arabia featured a highly dynamic genetic pool over time, since its successful settlement at ~60 thousand years ago (ka) during the out-of-Africa dispersal [16,24]. The Arabian Peninsula was exposed to several climate change episodes, with fluctuations between arid (leading to population contraction) and humid (population expansion) phases, which conditioned its role as a bridge connecting Africa with Eurasia [25,26].This bridge may have been limited, over long periods or in climatically unfavourable times, to three refuge areas: the Red Sea coastal plain; the Dhofar and Mahra Mountains and adjacent littoral zone in Yemen and Oman; and the emerged floodplain within the Persian Gulf basin [27]. In particular, the latter "Gulf Oasis" may have been fundamental for the survival during arid conditions of the ancient N(xR) mtDNA lineages coalescing at ~60 ka found in Arabia [24], most likely the relicts of the first migrants; the Gulf was also a preferential contact bridge with the Fertile Crescent.

Since these relict lineages are very minor, however, this signal for the settlement of Arabia during the successful out-of-Africa migration does not clarify if it was a continuous process lasting to the present day. The Pleistocene to Holocene continuity *versus* discontinuity debate has centred on how far the Arabian population was made up from the producers of the Levantine Pre-Pottery Neolithic B (PPNB)-related industry [28]. After rather sparse Late Palaeolithic settlement, the archaeological evidence suggests a significant increase in sites throughout Arabia dating from 9–8 ka [29], but it remains unclear if these were the result of newly arrived people [30] or locals who adopted the new food-producing technology [31]. The scarcity of secure stratigraphic reconstructions in the archaeology of the Peninsula has contributed to the uncertainty in dating the major demographic events. We have shown that some of the most frequent South Arabian mtDNA lineages (such as R0a) display signs of introduction and expansion in the post-glacial period [32], thus pre-dating the Neolithic, although the global contribution of this period to the total Arabian maternal gene pool remains to be evaluated.

The archaeological evidence is clearer regarding the remarkable maritime trade system that Arabia established with Africa, the Near East and India in the ninth to eighth millennia, probably the earliest worldwide [33]. The maritime traffic was intensified in mid-sixth millennium, with the appearance of the Pre-Dynastic Egyptian period, which dominated long-distance trade in the Red Sea [34], while in the Persian Gulf trade was established between communities in present-day Bahrain, the Oman Peninsula, the Indus Valley and Gujarat [35].This trade contributed to commercial, cultural, linguistic and genetic exchanges. In terms of language expansion in the region, by applying a Bayesian approach to Semitic lexical data, Kitchen et al. [36] concluded for a single entrance of early Ethio-Semitic languages in Africa, from southern Arabia, at around 2800 years ago, a period when South Arabia was influential in northern Ethiopia. A well-documented movement of people occurred through the Arab slave trade established between the 6th and 19th centuries AD [37], bringing African people (from Nubia to Zanzibar) into the Near East, Arabian Peninsula and even India and China. Estimates indicate that 2,400,000 African people were enslaved along the Red Sea and Indian Ocean routes [38], with a 2:1 female to male ratio [39]. This has also been proposed to explain the high levels of African L(xMN) lineages observed in Yemen [37,40], but other potential sources for sub-Saharan African (but also Indian and Southeast Asian) mtDNA lineages in Arabia may be the result of Hadrami men spending several generations in diaspora around the Indian Ocean rim and returning to their homeland with women taken from the diaspora [41]. Kivisild et al. [37] also detected a 12% frequency of haplogroup L6 in their Yemeni population sample from Kuwait, which is only being marginally observed in Ethiopia and almost absent elsewhere in Africa, and hypothesised that L6 originated from the successful out-of-Africa migration at ~60 ka.

However, the subsequent characterisation of other Arabian populations, including Yemen and Oman [42,43,44,45,46,47], did not reproduce the high frequency of this mtDNA lineage in South Arabia.

In this work, we use mtDNA to provide a detailed stratigraphic characterisation of key demographic events in Arabia since the first successful out-of-Africa migration ~60 ka. We performed mtDNA founder analysis for Arabia and neighbouring regions, aiming to ascertain and date the main dispersal episodes. The founder analysis was applied to the unbiased HVS-I database available for the region, and interpreted in the light of the more precise dating information gathered from whole-mtDNA sequences of informative haplogroups [24,32,47,48]. We also updated the phylogenetic trees of haplogroups J, T, L4 and L6, by performing 83 new whole-mtDNA sequences. We further tested our inferences from the HVS-I based founder analysis with a whole-mtDNA founder analysis using haplogroups J and T. The mtDNA information is put in perspective with results from genome-wide analyses of published data [3,23,49,50], focused for the first time on inferring the local Arabian population structure, which has been overlooked in the worldwide context of previous autosomal work.

Results/Discussion

Continuity of Pleistocene/Holocene settlement

Previous work has already provided genetic evidence for the exchange of lineages between the Near East and Arabia. This was confirmed with whole-mtDNA sequencing of the Eurasian macrohaplogroup N (including its branches X, I, W, N1a, N1b and some R lineages), which is dominant in Arabia, attaining a frequency of 66%–83% [24,32,47,48]. The obvious missing element in those studies was the whole-mtDNA sequencing of Arabian JT lineages, which we have performed here, providing a detailed phylogeographic analysis in Supplemental Material (outline topology in S1 and S2 Figs.; S1 Text). Following the pattern for the remaining N lineages, the frequency and diversity maps (S3, S4, S5, S7, S12, S13, S16 and S19 Figs.; S3 and S4 Tables) of JT lineages, displaying similarity across the Near East and Arabian Peninsula, as well as the many basal Arabian lineages (S8, S9, S10, S11, S14, S15, S17, S18, S20, S21, S22 and S23 Figs.), suggest that both regions were in close contact throughout the late Pleistocene and Holocene. Haplogroup J assumes a more important role in Arabia overall than haplogroup T, as testified by frequencies (between 7.7–20.6% and 3.2–10.2%, respectively) and the many star-like J sub-clades observed in Arabia, dating to ~6–7 ka. These expansions in haplogroup J are reflected in the BSP analysis (S6 Fig.), for which the main increase in effective size was between 8–12 ka in Arabia (S6A Fig.), after the expansion observed in the Near East around 11–15 ka (S6B Fig.). Haplogroup J also shows signs of having crossed into eastern Africa, particularly the sub-clade J1d1a1, necessarily after its emergence in Arabia at ~7.1 ka (S14 Fig.). Thus haplogroup JT indicates that demographic expansion in Southwest Asia was a continuous phenomenon from the Late Glacial period to the Neolithic period.

In order to dissect the apparent continuous genetic exchange between Arabia and the Near East since the late Pleistocene, we performed a founder analysis for all Eurasian haplogroups assuming the Near East, Iran and Pakistan as source and Arabia as sink (identified founders reported in S6 and S7 Tables). Fig. 1A displays the overall pattern, which seems to favour the periods around 1ka, 10 ka and 16 ka for migrations. Based on this information, we further imposed these dates as migration events to represent broadly, respectively, recent events, the Younger Dryas/Neolithic transition and the Late Glacial period. The results indicate that the Late Glacial period (Fig. 1B) was the most important migratory period, responsible for the introduction of 40–54% of the lineages (mainly belonging to the haplogroups K, U2, U3, U4, N1a1a, N1a1b, H5 and HV1; S24 and S25 Figs. and detailed description in S1 Text). At the

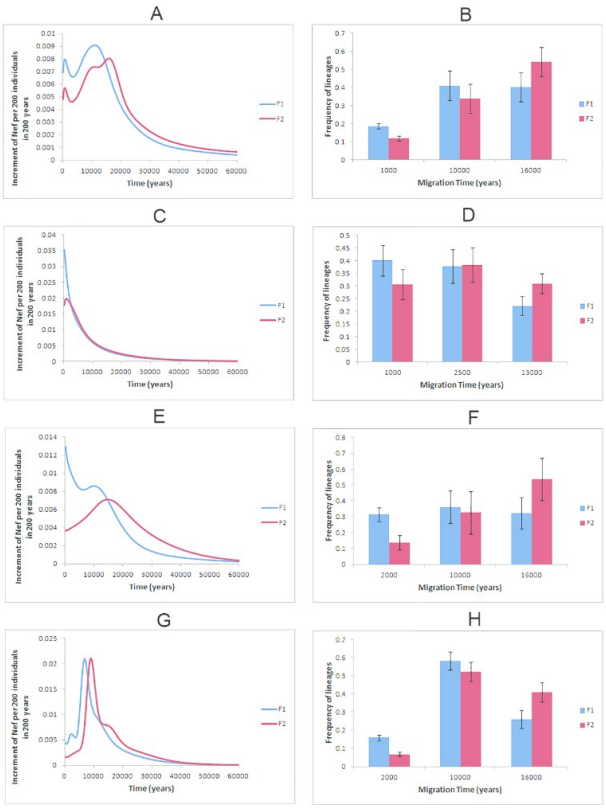

Fig 1. Founder analysis results. Probabilistic distribution of founder clusters across migration times, with time scanned at 200 year intervals from 0–60 ka, using *f1* (blue line) and *f2* criteria (red line), when considering putative migrations: (A) from the Near East, Iran and Pakistan to Arabia; (C) from Africa into Arabia plus the Near East and Iran; (E) Arabia plus the Near East and Iran into eastern Africa; (G) Arabia plus the Near East and Iran into North Africa; and probabilistic proportion of founder clusters considering different migration events, using *f1* (blue bar) and *f2* criteria (red bar), when considering putative migrations: (B) from the Near East, Iran and Pakistan to Arabia; (D) from African into Arabia plus the Near East and Iran; (F) Arabia plus the Near East and Iran into eastern Africa; (H) Arabia plus the Near East and Iran into North Africa.

doi:10.1371/journal.pone.0118625.g001

Younger Dryas/Neolithic boundary, 34–41% of lineages, mainly unclassified HV, R0a, J1b, T1a and M1 migrated to Arabia. The remaining 12–19% moved very recently, ~1 ka, and consists of derived lineages, (including J1d1a, K1, HV8 and N1a3). Although it is hard to discriminate clearly between the Near Eastern and Pakistan/Iranian influences, due to their largely shared mtDNA pool, the results suggest a higher Pakistan/Iranian impact in the east (41%) than in the west (25%) of Arabia for private founders, but just 14% and 11%, respectively, when considering the overall pool. This seems to indicate that the Pakistan/Iranian contribution was recent,

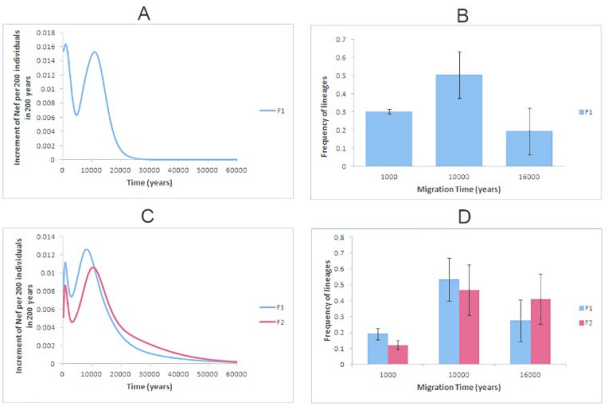

Fig 2. Founder analysis results on JT lineages. Probabilistic distribution of founder clusters across migration times, with time scanned at 200 year intervals from 0–60 ka, using *f1* (blue line) and *f2* criteria (red line), when considering putative migrations from the Near East, Iran and Pakistan to Arabia for (A) whole-mtDNA genomes or (C) HVS-I for haplogroups J and T; and probabilistic proportion of founder clusters considering different migration events, using *f1* (blue bar) and *f2* criteria (red bar), when considering putative migrations from the Near East, Iran and Pakistan to Arabia for (B) whole-mtDNA genomes or (D) HVS-I for haplogroups J and T.

doi:10.1371/journal.pone.0118625.g002

as the lineages introduced from this region did not reach high frequencies, and as expected its impact was higher in the eastern Arabian countries.

We next tested the robustness of the founder analysis by using whole-mtDNA genomes and HVS-I from haplogroups J and T alone (Fig. 2). The 17 whole-mtDNA founders identified (S8 Table) contributed to the overall pattern of migration displayed in Fig. 2A, which displays two main peaks, at 1 ka and 10 ka. When imposing the model of three migrations (Fig. 2B), 30% of JT lineages were introduced at 1ka, 50% at 10ka and 20% at 16ka. These results match closely the inferences based only on HVS-I information (Fig. 2C,D).

We should emphasize that no one-to-one correspondence of founder types between whole-mtDNA genomes and HVS-I can be expected, as there is no such precise correspondence between the whole-mtDNA and HVS-I trees, due in part to the differences in resolution but also no doubt to the small samples size at present for the whole-mtDNAs. We must also beware that other factors may also confound the analysis in particular circumstances. An extreme—but very unusual—instance is haplogroup J1d1a. Here, the HVS-I based founder analysis dates the founders to 1.0 ka, while the whole-mtDNA analysis indicates that it expanded in Arabia at least 6.1 ka. This discrepancy is due to 18 HVS-I sequences belonging to the root haplotype largely from central Saudi Arabia, an artefact of the sampling location (central Saudi Arabia is extremely arid and has had historically very low population size, with habitation restricted to oases, undoubtedly leading to severe genetic drift), while the remaining more diverse samples are from Yemen (as for most of the whole-mtDNAs). If the Saudi samples are disregarded, a ρ estimate for the founder age in Arabia increases to ∼6–7 ka, fitting more closely with the whole-mtDNA result. Allowing for such inevitable noise effects from the datasets, the similarity between the whole-mtDNA and HVS-I analyses is indeed striking, and we conclude that it is reasonable to infer that the picture suggested by the whole-population HVS-I founder analysis is not giving a very misleading impression of the dispersal history of the region.

Although it is not possible to date securely events as old as the ones occurring in the Pleistocene/Holocene transition based on genome-wide data alone, it is interesting to observe how the patterns of shared genome-wide ancestry support the inferences made for the mtDNA. All the Arabian populations form a close group with Near East populations in PC analysis (Fig. 3), with the first component explaining 44% of the diversity and partitioning populations along a west–east axis, and the second component explaining 8% and organising populations on a north–south axis. A few individuals in Arabian populations most probably had recent ancestry within Africa (especially for Yemen) or Pakistan (in the United Arab Emirates; UAE). Yemen shows the highest dispersion along the first axis, testifying again the higher African input in the closest country to the Horn of Africa. We confirmed the clustering of Yemeni Jews with Bedouin and Saudi Arabians, already identified previously [23], and probably indicating that they were less open to recent admixture with non-Arabian populations than their Yemeni Arab/Muslims neighbours.

The ADMIXTURE results indicate that K = 6 (Fig. 4 and Table 1; other K plots are displayed in S38 Fig.) is the number of clusters that best represents the population structure of the analysed populations. Here it is already possible to distinguish between a Southwest Asian/Caucasian and an Arabian/North African component; these two components have similar proportions of ~30% each in Yemen and UAE, but the Arabian/North African proportion increases to 52–60% in Saudi and Bedouin. In Near Eastern populations, correspondingly, the Southwest Asian/Caucasian component rises to ~50% and the Arabian/North African cluster decreases to ~20–30%, even in Palestinians (similar to the Samaritans and some of the

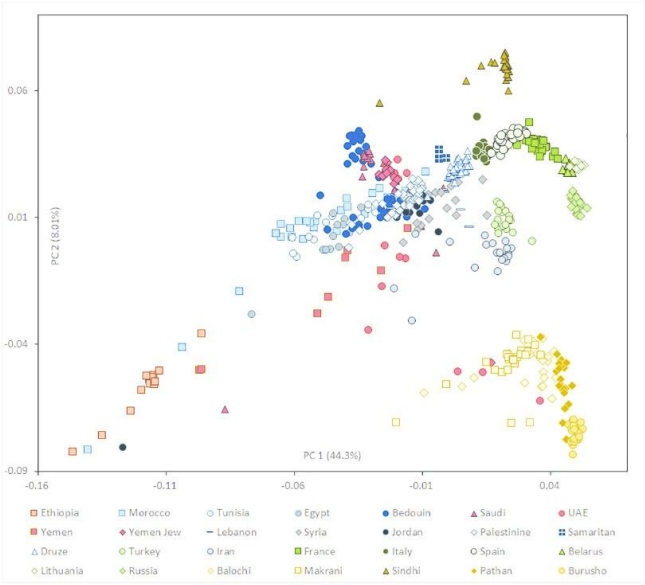

Fig 3. PCA results. Scatter plot of individuals, showing the first two principal components. Each symbol corresponds to one individual and the colour indicates the region of origin.

doi:10.1371/journal.pone.0118625.g003

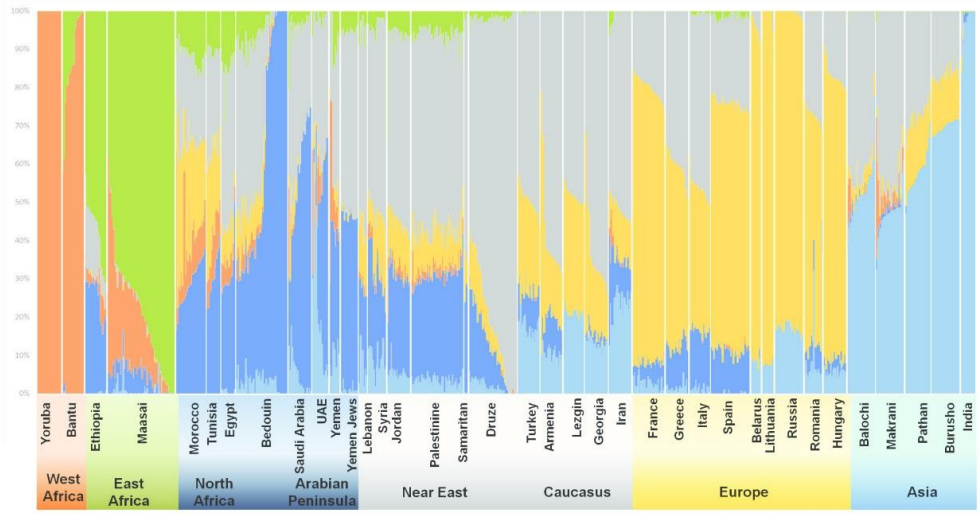

Fig 4. ADMIXTURE results. Population structure inferred by ADMIXTURE analysis. Each individual is represented by a vertical (100%) stacked column of genetic components proportions shown in colour for K = 6.

doi:10.1371/journal.pone.0118625.g004

Druze), highlighting their primarily indigenous origin, with the most extreme values for the Druze, carrying the Southwest Asian/Caucasian component at ~80%.

European background is higher in Near Eastern populations (around 9–15%) than in Arabia (1.5–5%) while the African ancestry is ~25% in Yemen, and then 4–8% in all Arabian and Near East populations except in Samaritans and Druze, with 0–2%. The UAE has a substantial pool from South Asia (21%) similar to the proportion displayed in Iran (24%), which falls to below 10% in all other Arabian and Near Eastern populations, except Turkey (18%).

ADMIXTURE allows us to calculate F_{ST} values between the components in order to quantify their similarity (Fig. 5A). For K = 6, Arabia showed a lower distance from the Near East (0.046), than from Europe (0.052), eastern Africa (0.098) and finally western Africa (0.140). Arabia and the Near East have similar genetic distances from eastern African (0.098 and 0.097, respectively), double that of the value between western and eastern Africa (0.046). When evaluating F_{ST} values in pairwise comparisons between Arabian and Near Eastern populations (Fig. 5B), we see that F_{ST} values are higher between Yemen and all other populations (and also for comparisons with Samaritans, but these results may be biased by low sample size). The UAE is closer to Jordan, Syria and Lebanon than Saudi Arabia is; while Saudi are closer to Palestinians, Druze and Samaritans than UAE. Thus, F_{ST} values support lower or similar genetic distances between UAE and Near Eastern populations as between Saudi and Near Eastern populations, while Yemen is clearly more divergent.

Exchanges across the Red Sea—from Africa into Arabia

Founder analysis of the dispersal of sub-Saharan lineages from Africa into Arabia plus the Near East and Iran (both regions have to be considered together due to the relatively low number of L(xMN) sequences) showed a predominant migration peak at 0–0.8 ka (Fig. 1C). When

Table 1. Estimates of admixture proportions (%) and date of admixture (in generations) calculated in ROLLOFF when using western (Yoruba) and eastern (Maasai) African and Italians + Spanish as ancestral populations.

Population	Sample Size	Western African ancestry proportion (%) ± standard error	Eastern African ancestry proportion (%) ± standard error	Southwest Asian/ Caucasian ancestry proportion (%) ± standard error	Arabian/ North African ancestry proportion (%) ± standard error	European ancestry proportion (%) ± standard error	South Asian ancestry proportion (%) ± standard error	Estimated date of admixture using ROLLOFF using Western African parental population	Estimated date of admixture using ROLLOFF using Eastern African parental population
Yemen	9*	16.935 ± 15.960	7.747 ± 5.333	30.777 ± 9.896	32.398 ± 6.030	3.217 ± 2.77	8.926 ± 3.727	21.019 ± 7.450	11.556 ± 3.878
Saudi Arabia	20	1.694 ± 5.223	4.033 ± 4.235	34.227 ± 8.955	52.479 ± 18.957	2.722 ± 3.879	4.844. ± 4.975	30.762 ± 4.907	25.430 ± 3.011
Yemen Jews	15	0.001 ± 0.000	5.105 ± 0.826	47.542 ± 1.525	45.693 ± 1.598	0.565 ± 0.699	1.094 ± 1.187	n/a	n/a
UAE	14	6.408 ± 9.118	1.817 ± 2.014	34.432 ± 4.312	34.378± 21.632	1.689 ± 1.931	21.276 ± 17.660	8.900 ± 1.642	8.923 ± 1.795
Bedouin	45	2.005 ± 2.213	4.692 ± 4.246	24.903 ± 19.909	60.057± 30.707	5.400 ± 4.700	2.944 ± 2.285	37.546 ± 3.104	27.734 ± 1.532
Lebanon	7	1.243 ± 4.854	4.670 ± 3.148	51.547 ± 2.519	21.092 ± 4.062	14.543 ± 2.791	6.905 ± 4.854	n/a	n/a
Syria	16	1.586 ± 1.451	3.413 ± 1.952	49.742 ± 4.880	23.260 ± 5.283	12.864 ± 4.532	9.135 ± 3.387	37.334 ± 4.365	26.181 ± 4.428
Jordan	20	3.205 ± 5.629	7.289 ± 6.404	47.833 ± 7.442	25.055 ± 3.209	11.171 ± 2.436	5.447 ± 2.169	32.871± 4.106	29.470 ± 3.671
Samaritan	3	0.001 ± 0.000	0.190 ± 0.777	63.029 ± 2.282	26.358 ± 2.709	8.946 ± 4.104	0.475 ± 0.496	n/a	n/a
Druze	42	0.178 ± 0.365	1.869 ± 1.082	80.100 ± 14.498	9.919 ± 7.730	6.123 ± 5.089	1.812 ± 1.664	n/a	n/a
Palestinian	46	2.222 ± 1.760	6.119 ± 2.147	51.538 ± 4.397	27.396 ± 2.153	9.153 ± 1.826	3.572 ± 1.302	29.008 ± 2.194	11.556 ± 3.878
Iran	20	1.701 ± 3.196	1.022 ± 1.818	50.678 ± 4.259	11.850 ± 5.614	11.135 ± 2.916	23.614 ± 3.944	n/a	n/a
Turkey	19	0.069 ± 0.029	0.194 ± 0.312	49.188 ± 3.258	8.993 ± 2.904	23.798 ± 3.503	17.758 ± 2.504	n/a	n/a
Ethiopia	19	3.911 ± 3.047	58.139 ± 8.479	12.146 ± 5.638	25.469 ± 5.495	0.179 ± 0.442	0.157 ± 0.297	93.223± 9.678	n/a
Maasai	19	15.808 ± 12.911	78.060 ± 15.009	0.412 ± 0.911	4.120 ± 3.043	0.096 ± 0.315	0.736 ± 1.858	47.007± 2.933	n/a
Egypt	12	5.553 ± 1.553	15.117 ± 4.878	39.826 ± 5.130	30.499 ± 6.343	8.380 ± 2.245	0.624 ± 0.630	30.034± 3.233	22.766 ± 2.890
Morocco	25	12.199 ± 10.473	12.066 ± 2.951	21.360 ± 4.827	28.872 ± 5.736	25.502 ± 7.971	0.001 ± 0.000	n/a	n/a
Tunisia	12	9.815 ± 2.927	10.437 ± 1.212	26.002 ± 4.057	30.991 + 6.178	22.754 ± 5.354	0.001 ± 0.000	n/a	n/a

N/A—not assigned.

* By eliminating one individual with a high level of African ancestry.

doi:10.1371/journal.pone.0118625.t001

checking these founders (S9 and S10 Tables), we see that most of them display clearly young ages, but several have ages ~13 ka (S15 Table). So, we tested a model based on three periods of migration (Fig. 1D), and their impact was: 31–40% for 1 ka (middle of Arab slave trade, 6th–

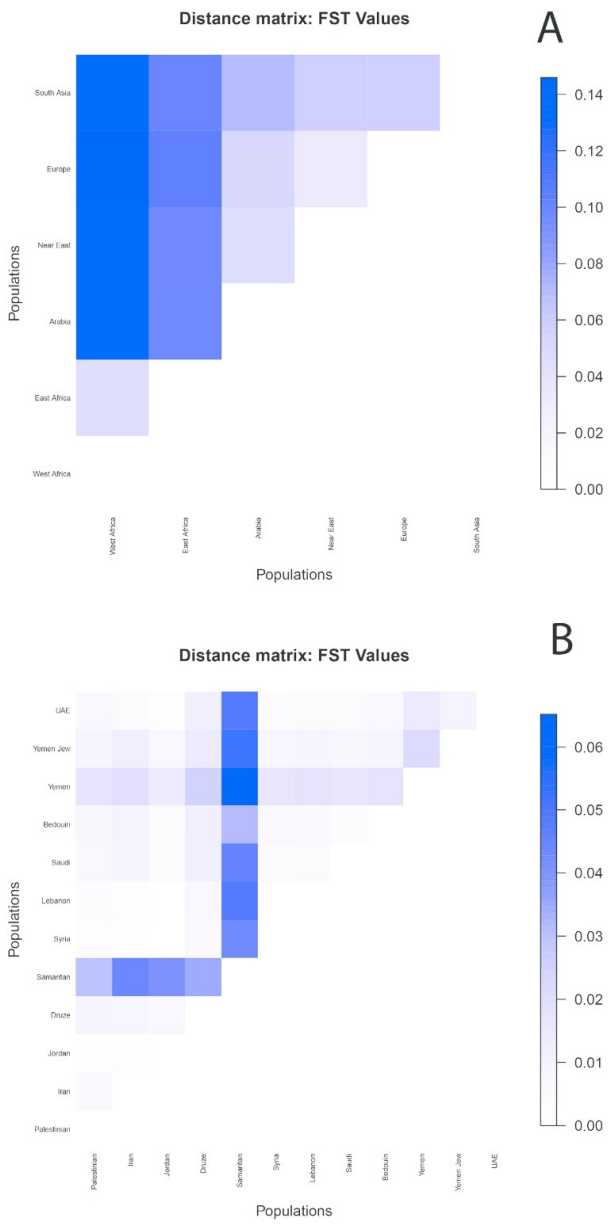

Fig 5. Matrices of F_{ST} distances. Matrices of F_{ST} values between ADMIXTURE components (A) and Arabian and Near Eastern populations (B).

doi:10.1371/journal.pone.0118625.g005

19th centuries); 38% for 2.5 ka (Arabian dominance of the Red Sea trade routes); and 22–31% for 13ka (close to the Younger Dryas). As the great majority of lineages migrated in the two very recent putative events, at similar ages, this contributes to the dominant young peak in Fig. 1C, while the approximately one-third of sequences that were introduced later is responsible for the long tail of the curve (instead of a sharper peak). No clear pattern of association between haplogroup and event was observable, probably reflecting high levels of heterogeneity in the source (S32 and S33 Figs. and detailed description in S1 Text). Thus, the Arabian maritime dominance and slave trade (from very recently, back until ~2.5 ka) were the main contributors (~69–78%) to the African ancestry into Arabia, Near East and Iran, but the entrance seems to have been initiated as early as the end of the Pleistocene. Clearly, no lineages could be assigned to the out-of-Africa migration event.

In order to provide more information to the issue of possible relicts of the out-of-Africa migration, we further investigated two relatively rare African haplogroups (L4 and L6), phylogenetically close to L3, by whole-mtDNA sequencing (outline topology in S26 Fig. and detailed topology in S28, S29 and S30 Figs.; S1 Text). L4 is more frequent nowadays in eastern Africa followed by the Near East (S27A Fig.; S5 Table). The whole-mtDNA-based date points to an origin at ~87 ka, predating the out-of-Africa dispersal (as well as its sub-clade, L4b, dating to ~86 ka). So, in theory, this sister haplogroup of L3 could have crossed into Arabia along with L3 during the initial out-of-Africa movement. Phylogenetically, however, the few Arabian L4 lineages are derived, supporting an explanation in more recent exchange networks between eastern Africa and Arabia for their dispersal, concordant with the recent signs of population growth detected for L4 in BSPs (S31A Fig.; and dominating also S31B Fig.; S14 Table). L6, at similarly low frequencies in Yemen and eastern Africa (S27B Fig.), dates to 23.1 [15.8–30.5] ka, and is likely to have migrated from eastern Africa into Arabia after that period, most probably very recently as testified by a very derived L6a sub-clade observed in three Yemenis (sharing the same lineage).

The genome-wide analyses performed here on the available data from Arabian populations provide estimates of African admixture, with disentanglement between western and eastern African gene pool contributions (Table 1). The eastern African background is around 4.0% in Saudi and Bedouin, ~7.7% in Yemen (although Yemen Jews have a lower admixture of 5.1%), and 1.8% in UAE; this input decreases beyond Jordan, and is negligible in Samaritans, Druze, Turks and Iranians. The western African component also varies between 2.0 and 6.4%, except for Yemen (16.9%) where it has likely been inflated due to indirect recent migration (the Bantu component which is present in many eastern African populations). The ROLLOFF estimates for the event of admixture were 8–27 generations ago when using eastern Africa as parental population, and 8–37 generations using a western African source.

Both date estimates are compatible with the Arab slave trade, which operated between the 6th and 19th centuries AD, mainly from eastern Africa (from Nubia to Zanzibar), although many of these populations bear a significant western African component (as shown in Fig. 4). These values are in agreement with the estimates of Moorjani et al. [1] for Levantine groups, showing a 4–15% African ancestry and about 32 generations ago for the event of admixture, interpreted as consistent with close political, economic, and cultural links with Egypt in the late Middle Ages. They also estimated 72 generations ago for the event leading to 3–5% sub-Saharan ancestry in diverse Jewish populations, arguing that this reflecting descent of these groups from a common ancestral population that already had some African ancestry prior to the Jewish Diaspora.

Hodgson et al. [7] focused on the back-to-Africa migration in the Horn of Africa, and obtained ages from 2.2–4.7 ka for the admixture event when using the ROLLOFF and ALDER methods. The authors relied on other approaches in order to evaluate the hypothesis of two or

more distinct episodes of non-African admixture in the Horn of Africa: they identified a non-African Ethio-Somali component in eastern African populations in the ADMIXTURE analysis for which F_{ST}-based dating methods indicated an age of divergence from North African/Arabian populations of 23–25 ka, leading to a possible window of migration pre-LGM. These results fit well with the conclusions we reached in this study through the analysis of the maternal mtDNA pool.

Exchanges across the Red Sea—from Arabia into Africa

The Bab-el-Mandab strait and the Red Sea were also important for dispersal in the opposite direction, the "back-to-Africa" migrations. Founder analysis (Fig. 1E; S11 and S12 Tables) led to the identification of peaks of migration at ~10–15 ka. Given these results, we inferred two main migration events, at ~10 ka (representing the Neolithic and beginning of maritime trade) and at ~16ka (Late Glacial period), as well as an episode at ~2 ka which could represent recent times (specifically, Arabian dominance of the Red Sea routes). The proportions (Fig. 1F) for migration contributed by these events were: 14–31% at ~2 ka (for N1, R0a, T, J, K and X); 33–36% at ~10ka (U6a1a, J1d1a, M1 and R0a); and 33%–54% at ~16ka (M1 and HV1). A detailed analysis of these haplogroup distributions in the migration events is provided in S1 Text, S34 and S35 Figs.

Interpreting these results in the light of available whole-mtDNA sequences, only the introduction of N1 seems younger than expected, most probably due to lack of HVS-I resolution for this haplogroup. Two main founders (comprising haplogroups N1a and I) are at the root of N1 sub-clades (dating to 15.9 and 21.8 ka, respectively). Another founder in N1a could be placed in the sub-clade identified in the whole-mtDNA sequencing from Somalia reported by Fernandes et al. [24], bearing the substitution at position 16213; but the HVS-I data show that this is more frequent in Africa (seven individuals) than in Arabia (one individual), so this Arabian individual may be a recent introduction into Arabia of an N1a sub-clade that had evolved within Africa (dating to 0.9 ka [24]).

The phylogenetic analyses for N(xR) lineages performed by Fernandes et al. [24] also provided insights into back-to-Africa movements, evidently at various time periods. Some lineages (I, N1a and N1f) displayed deep branches in eastern Africa, a sign of introduction in Africa which could have begun as early as ~40 ka (the upper bound defined by the TMRCA of the founder clades) and extending till ~15 ka (the lower bound defined by the TMRCA of the derived African clades). The migration of J1d1a lineages into eastern Africa in the Neolithic period is confirmed in the whole-mtDNA sequencing (S14 Fig.) and complemented by the frequency interpolation and founder analysis (S13 Fig.) performed here.

From the genome-wide results, we can infer this back-to-Africa migration was considerable, leading to a proportion of 12% of Near Eastern and 26% Arabian ancestry in Ethiopia (Table 1). The ROLLOFF estimate for the date of admixture was 93 generations ago—twice as old as the time of African admixture in Arabia and Near East. For comparison, in the Maasai from Kenya and Tanzania, the Eurasian component is an order of magnitude lower (4.5%), and the time of admixture is 47 generations, reflecting most probably later admixture events.

The parallel introduction of Eurasian lineages from the Near East, Iran and Arabia into North Africa through the Sinai Peninsula revealed two well-defined peaks (Fig. 1G) at ~2.4 ka and 6.8 ka with the *f1* criterion, and two peaks at ~9.0 ka and ~12.4 ka when using the *f2* criterion. This seems to point to a significant role for dispersal in the Neolithic period, consistent with results obtained for the North African MSY pool, interpreted as suggesting a large Neolithic origin [51]. A major Neolithic impact is supported when imposing periods for the migration of founders (Fig. 1H), leading to: 7–16% at ~2 ka, mainly HV1 and other undefined HV

lineages, M1 and U (U6a1, K1a1); 52–58% at ~10 ka for most of HV, U (U5b, U5 and K), T (some T2c1 and T2b), J (J1d1a, J2a2b and other undefined J), and X; and 26%–41% at ~16 ka for some HV, T (T1a, T2) and U (U3, U3a, U5b1b, U5a, U6a) lineages (S1 Text, S36 and S37 Figs.). It seems likely that some JT lineages, especially T ones, were introduced into Northeast Africa before the Neolithic, following Late Glacial population expansions in the Near East/Arabia. Then, locally they could have been involved in population expansions in the Neolithic period, leading to signs of autochthonous founder effects, such as the one detected in the El-Hayez oasis (400 km southwest of Cairo) for sub-haplogroup T1a2a [52].

The link between U6 and M1 and the settlement of North Africa from the Near East at ~45 ka advanced previously [53,54] was recently put into question [55] because their sub-clades do not all seem to display the same history: U6a is ~10 ka older than M1a and M1b, and sub-clades of the former coalesce before or around the LGM while sub-clades of the latter date to the post-LGM. In our founder analysis for North Africa, a strong Late Glacial signal was detected for U6.

At the genome-wide level, Egypt is quite similar to its Levantine neighbours, displaying a mainly Near Eastern (39.8%) and Arabian/North African (30.5%) background, with slightly higher western (5.6%) and eastern (15.1%) African proportions, and lower European (8.4%) and South Asian (0.6%) proportions. The ROLLOFF estimate for admixture in Egypt (using Africans and Europeans as ancestral populations) was 30 generations, predictably young due to continuous gene flow between the two regions. Morocco and Tunisia presented similar western (9.8–12.2%) and eastern African (10.4–12.1%) components and roughly twice the magnitude for each of the European (22.8–25.5%), Near Eastern (21.4–26.0%) and Arabian (28.9–31.0%) pools. Again these young dates show that simple genome-wide dating approaches based on linkage disequilibrium decay must be applied cautiously in complex scenarios of several migrations occurring over a long span of time, such as the ones which took place across the Red Sea, North Africa [56] and Iberia [57].

Conclusions

The detailed evaluation of the Arabian and neighbouring mtDNA pools has allowed us to establish a genetic stratigraphy of Arabia's maternal line of descent, testifying to the pivotal role of the Peninsula at the crossroads between Africa and Eurasia. The successful out-of-Africa migration led to continuous settlement of parts of the Peninsula, most probably centred on the Gulf Oasis, which likely functioned as the cradle for the emergence of the haplogroup N lineages. No haplogroup L(xMN) relicts of this migration into Arabia are detected in mtDNA founder analysis and we have confirmed their absence by whole-mtDNA sequencing of lineages from L3 [16] and its sister clades L4 and L6.

Although it is likely that the Gulf Oasis region eventually formed part of an extended source region together with the Near East, if we assume that the Near East was the main source population for current Arabian diversity, the Late Glacial period was responsible for the introduction of 40–54% of lineages, the Younger Dryas/Neolithic for 34–41%, and recent times (at 1.0 ka) for the remaining 12–19%. The Neolithic in Arabia was more characterised by the expansion in effective size of local haplogroup N lineages, mostly within R0a and J, than by the entrance of new lineages. Arabia, together with the Near East and Iran, was involved in the "back-to-Africa" migration of Eurasian lineages, beginning in the Pleistocene but becoming more significant with the establishment of maritime commercial routes. The Late Glacial period was more important for bringing Eurasian lineages into eastern Africa, probably reflecting the higher impact of this period in the expansion of Arabian populations, while the Neolithic, especially linked to the Near East, affected to a greater extent the dispersals towards North

Africa. The biparental genome averaged the African input to 6–25% of the Arabian pool, concordant with the 35% female and 0% male inputs estimated from uniparental systems. ROLL-OFF dating of admixture events across the Red Sea suggested recent ages of 8–37 generations for the African input into Arabia, 93 generations for the Arabian/Near Eastern input into eastern Africa and 30 generations for North Africa.

We conclude by emphasising that different parts of the genome of an admixed population often tell different stories—so the strategy must involve independent evaluation of (large) linked blocks. This is precisely what we do when analysing the diverse mtDNA lineages found in a population, but because mtDNA is a single linked locus, the different stories then emerge from the different lineages, carried by different individuals within a population. Probably, regions of the nuclear genome with a low recombination rate will allow estimation of older events, as soon as more complete nuclear genomes are available from more populations, overcoming the limits of molecular resolution of current genome-wide SNPs.

Materials and Methods

Samples for whole-mtDNA sequencing and statistical comparisons

We previously characterised the mtDNA diversity in populations from eastern Africa [16], the Arabian Peninsula [42,46,47], and the African Sahel [58], by sequencing the hypervariable segments I and in some cases II (HVS-I and HVS-II) using a procedure described previously [59]. This information was used to assign mtDNA sequences to haplogroups, following the most up-to-date phylogenetic evidence, reported on the PhyloTree website [60],checking the classification against the output of the Haplogrep software [61]. We then selected 26 UAE and 31 Yemen samples belonging to haplogroups J and T, and some belonging to haplogroups L4 and L6 for whole-mtDNA sequencing, amounting into a total of 26 (L4: 1 Burkina Faso, 2 Chad, 2 Dubai, 4 Ethiopia, 2 Kenya, 1 Niger, 1 Nigeria, 1 Nubia, 5 Somalia and Sudan; L6: 2 Ethiopia, 1 Kenya and 2 Somalia) (S1 Table).

We followed the methodology and quality control procedures of Pereira et al. [62], and mutations were scored relative to the revised Cambridge reference sequence [63]. The sequences obtained are reported in S1 Table and have been deposited in GenBank (accession numbers KP316996-KP317078).

For the whole-mtDNA analyses (S1 and S2 Tables), we used a total of 1779 samples of JT whole-mtDNA sequences (57 new, 1722 published) and 57 L4/L6 sequences (26 new, 31 published) in the reconstruction of their phylogenetic trees. We constructed a database of HVS-I and HVS-II sets from African, Arabian, European, Near Eastern, Iranian and Pakistani populations, amounting to 42,485 sequences, for founder analysis; these data are summarised in S6, S7, S8, S9, S10, S11 and S12 Tables. By the Arabian Peninsula, we assumed the territory covered by present-day Oman, UAE (which together we sometimes identified as eastern Arabia), Saudi Arabia and Yemen (western Arabia) countries. In the Near East, we included Iraq, Jordan, Israel/Palestine, Turkey, Lebanon and Syria.

This study obtained ethical approval from the Ethics Committee of the University of Porto, Portugal (11/CEUP/2011). Written informed consent was obtained from all sampled individuals, except from illiterate people who provided oral consent and a fingerprint instead of signature. The Ethics Committee approved this procedure.

Statistical analyses of mtDNA data

For the phylogenetic reconstructions, preliminary reduced-median network analyses [64] led to a suggested branching order for the trees, which we then constructed most parsimoniously by hand. We used the mtDNA-GeneSyn software [65] to convert files.

In order to estimate the time to the most recent common ancestor (TMRCA) for specific clades in the phylogeny, we used the ρ statistic [18] and maximum likelihood (ML). We used ρ (the mean sequence divergence from the inferred ancestral haplotype of the clade in question) with a mutation rate estimate for the whole-mtDNA sequence of one substitution in every 3624 years, correcting for purifying selection, and a synonymous mutation rate of one substitution in every 7884 years [66]. Standard errors were estimated as before [67]. We obtained the ML estimates of branch lengths using PAML 3.13 [68], assuming the HKY85 mutation model with gamma-distributed rates (approximated by a discrete distribution with 32 categories). We converted mutational distance in ML to time using the same whole-mtDNA genome clock.

In order to investigate the population demography associated with the different haplogroups analyzed (J/T and L4/L6), we obtained Bayesian skyline plots (BSPs) [69] from BEAST 1.4.6 [70] for a total of 1720 and 57 (J/T and L4/L6, respectively) whole-mtDNA sequences with a relaxed molecular clock (lognormal in distribution across branches and uncorrelated between them) and the HKY model of nucleotide substitutions with gamma-distributed rates (10 gamma categories). BSPs estimate the effective population size through time using random sequences from a given population, but have also proved effective with individual haplogroups data [71]. For this analysis, we used a mutation rate of 2.6129×10^{-5}, previously calibrated using internal calibration points within the L3 phylogeny [16]. BEAST uses a Markov-chain Monte-Carlo (MCMC) approach to sample from the posterior distributions of model parameters (branching times in the tree and substitution rates). Specifically, we ran 100,000,000 iterations, with samples drawn every 10,000 MCMC steps, after a discarded burn-in of 10,000,000 steps. We checked for convergence to the stationary distribution and sufficient sampling by inspection of posterior samples. We visualized the Bayesian skyline plots (BSPs) with Tracer v1.3 [69]. We used a generation time of 25 years and forced the larger haplogroups to be monophyletic in the analysis: MCMC updates which violated this assumption were rejected. In order to perform a systematic comparison and description of the increment periods in the effective population size of the BSP, we calculated a rate of population size change through time.

To visualize the geographical distribution of haplogroups J, T, L4 and L6, we constructed interpolation maps using the "Spatial Analyst Extension" of ArcView version 3.2 (www.esri.com/software/arcview/). We used the "Inverse Distance Weighted" (IDW) option with a power of two for the interpolation of the surface. IDW assumes that each input point has a local influence that decreases with distance. The geographic location used is the centre of the distribution area from which the individual samples of each population were collected. The data used are listed in S3, S4 and S5 Tables.

In order to estimate the times of migrations into and from the Arabian Peninsula, we employed founder analysis [15]. This method assumes a strict division between assumed source and sink populations and two criteria (*f1* and *f2*) for identifying founder sequences to partly allow for homoplasy and back migrations, by ensuring that sequence matches are not at the tips of the source phylogeny. Founders must have at least one (*f1*) or two (*f2*) derived branches in the source population. The first step is to reconstruct, haplogroup by haplogroup, the HVS-I networks in the range 16051–16400 bp of the reference sequence [63]; we then identified founders and descendants using an in-house computer tool [72]; and finally we estimated the age of the migration of each founder using the ρ statistic [18], assuming an HVS-I mutation rate of one mutation every 16,677 years [66].

Four paths of migration were tested: (1) from Africa into Arabia plus the Near East and Iran (identified through the L(xMN) haplogroups); (2) from the Near East, Iran and Pakistan into the Arabian Peninsula (N haplogroups); (3) from Arabia plus Near East and Iran into eastern Africa (N and M1 haplogroups); and (4) from Arabia plus Near East and Iran into North Africa (N and M1 haplogroups). We included Pakistan in path (2) as we were also interested in

inferring the more eastern contribution into the Arabian Peninsula. In order to assess the error in the Bayesian partitioning across the different migration times realistically, we calculated an effective number of samples for each founder cluster. This was obtained by multiplying the number of samples for each founder cluster by a ratio of the variance assuming a star-like network and the variance calculated as in Saillard et al. [67].

We scanned the distribution of founder ages for each region, defining equally spaced 200-year intervals for each migration from 0–70 ka. For each case, we also investigated the proportion of introduction of lineages during putative migrations occurring in certain periods of time. We selected these migration events by combining three distinct lines of evidence: the peaks detected in the founder analysis; historical/archaeological evidence; and dates from whole-mtDNA sequences belonging to informative haplogroups in the region (such as R0a, JT, N1, N2, I, L3 and L4/L6). We represented the probabilistic proportions of introduction for each lineage at each of the putative migration periods in graphs resembling the images from the STRUCTURE analysis.

In order to further validate the HVS-I founder analysis into Arabia we compared it with the results obtained from a founder analysis using whole-mtDNA genomes belonging to haplogroups J and T. We only used an *f1* criterion (since the sampling from the source was too scarce to allow an *f2* criterion) and we detected 17 founders (S8 Table). The assumptions of the founder method do not allow the use of a time-dependent clock. Therefore, given the relatively small difference between the mutation rate for time zero (average 2562 years for a mutation to happen) and the mutation rate for the oldest founder (average 2667 years for a mutation to happen) we used the intermediate value (2614 years for a mutation to happen) as an estimate for the overall range. As with the HVS-I founder analysis, we performed a preliminary scan analysis and estimated relative contributions of JT lineages in a three-migration model.

Genome-wide database

We assembled genome-wide data for 790 samples from eight geographic groups (sub-Saharan Africa, North Africa, Arabian Peninsula, Near East, Iran, Europe, Caucasus and South Asia) from previously published data sets (S13 Table). The samples from Behar et al. [23] were genotyped using Illumina the 610K and 660K bead arrays, while those from Li et al. [49] were screened with Illumina 650K bead arrays, and those from Hellenthal et al. [3] with Illumina 660K bead arrays. We obtained the genotypes from Maasai, an ethnic group located in Kenya, from the HapMap phase III release (http://hapmap.ncbi.nlm.nih.gov/). We used PLINK 1.05 [73] to check that individuals and SNPs had a genotyping success of 97%. We used a Python in-house script to merge genotypes from the various chips and ended up with a total of 309,474 common autosomal single nucleotide polymorphisms (SNPs). We pruned the full dataset for linkage disequilibrium (LD), removing SNPs in strong LD ($r^2 > 0.4$) with nearby markers in a window of 50 SNPs (advanced by 10 SNPs each time); a total of 215,286 SNPs remained for further analyses.

Genome-wide statistical analyses

We analysed the 790 samples with the ADMIXTURE software [74] which provides a maximum likelihood estimation of the population structure. We tested several numbers of clusters or ancestral populations, K (from three to six), with termination criteria for independent runs for each K value established when the log-likelihood increased by less than 10^{-4} between iterations. We performed across-validation to check the K value with the lowest cross-validation error, which would represent the most accurate modelling choice.

We carried out the principal component (PC) analysis, which infers worldwide axes of human genetic variation from the allele frequencies of various populations, using the *smartpca* tool, available in the EIGENSOFT package [75]. We evaluated the statistical significance of each PC through the Tracy-Widom statistics, computed at the EIGENSOFT tool *twstats*. As we were focused in Arabia, we did not include all populations in the analysis, especially the western African ones, in order to maximise the resolution.

To estimate the ages of putative admixture events in populations displaying statistical evidence of admixture, we used the ROLLOFF method [1] implemented in the ADMIXTOOLS software package [8]. This method is based on the decay of admixture LD in the target population, performing a local ancestry inference. We ran the ROLLOFF method for Arabia and some Near Eastern populations, using the unpruned set, with Maasai individuals (from the HapMap dataset, selected after the ADMIXTURE analysis, as the ones displaying >80% eastern African ancestry) and Italy plus Spain (extracted from 1000 Genomes database; http://browser.1000genomes.org/index.html) as ancestral populations. We also performed this analysis by replacing Maasai by Yoruba, from western Africa, to check for the influence of the selected African ancestral population, and as some eastern African populations also have a high western African component (such as Luhya in Webuye, Kenya, in the 1000 Genomes database).

We plotted the correlation between SNPs as a function of genetic distance for all chromosomes. Ages (in number of generations) were estimated by fitting an exponential distribution to the decay of these correlation coefficients. The estimated age (in number of generations) for the admixture event is the average of dates for all chromosomes. The F_{ST} values between pairs of ADMIXTURE components (K = 6) were estimated using ADMIXTURE, while the ones between pairs of populations were performed using vcf tools (http://vcftools.sourceforge.net/).

Supporting Information

S1 Fig. Schematic tree of haplogroup J. Ages (in ka) indicated are maximum likelihood estimates obtained for the whole-mtDNA genome.
(TIF)

S2 Fig. Schematic tree of haplogroup T. Ages (in ka) indicated are maximum likelihood estimates obtained for the whole-mtDNA genome.
(TIF)

S3 Fig. Frequency maps based on HVS-I data for haplogroups J (A) and T (B).
(TIF)

S4 Fig. Distribution maps for haplogroup J for the diversity measures π (A) and ρ (B) based on HVS-I data.
(TIF)

S5 Fig. Distribution maps for haplogroup T for the diversity measures π (A) and ρ (B) based on HVS-I data.
(TIF)

S6 Fig. Bayesian skyline plot indicating hypothetical effective population size through time based on data from haplogroup J of Arabia (A) and Near East (B) and from haplogroup T of Arabia (C) and Near East (D).
(TIF)

S7 Fig. Frequency maps based on HVS-I data for haplogroups J1b.
(TIF)

S8 Fig. Phylogenetic tree of haplogroup J1b. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; insertions are indicated by a dot followed by the number of repetitions and the nucleotide position; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S9 Fig. Phylogenetic tree of haplogroup J1b1. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S10 Fig. Phylogenetic tree of haplogroup J1b1a. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S11 Fig. Phylogenetic tree of haplogroup J1b2. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S12 Fig. Frequency maps based on HVS-I data for lineages within haplogroup J defined by the transition at 16193, which mainly corresponds to haplogroup J1d, but can also include haplogroup J2d.
(TIF)

S13 Fig. Frequency maps based on HVS-I data for the sub-haplogroup J1d1a.
(TIF)

S14 Fig. Phylogenetic tree of haplogroup J1d1. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; deletions are indicated "d"; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S15 Fig. Phylogenetic tree of haplogroup J1d2. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; insertions are indicated by a dot followed by the number of repetition and the nucleotide position; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S16 Fig. Frequency maps based on HVS-I data for haplogroup J2.
(TIF)

S17 Fig. Phylogenetic tree of haplogroup J2a2. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S18 Fig. Phylogenetic tree of haplogroup J2a2a. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; insertions are indicated by a dot followed by the number of repetition and the nucleotide position; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S19 Fig. Frequency maps based on HVS-I data for the haplogroup J2a2b.
(TIF)

S20 Fig. Phylogenetic tree of haplogroup T1a. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; insertions are indicated by a dot followed by the number of repetition and the nucleotide position; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other

coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S21 Fig. Phylogenetic tree of haplogroup T2a1. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; insertions are indicated by a dot followed by the number of repetition and the nucleotide position; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S22 Fig. Phylogenetic tree of haplogroup T2c. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; insertions are indicated by a dot followed by the number of repetition and the nucleotide position; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S23 Fig. Phylogenetic tree of haplogroups T2i and T2g. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; insertions are indicated by a dot followed by the number of repetition and the nucleotide position; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S24 Fig. Probabilistic proportion of founder clusters considering three migration periods (1.0, 10.0 and 16.0 ka), using the *f1* criterion and by assuming a Near East, Iran and Pakistan source for migrations into Arabian Peninsula. The haplogroup affiliations of the founders are indicated in the bottom.
(TIF)

S25 Fig. Probabilistic proportion of founder clusters considering three migration periods (1.0, 10.0 and 16.0 ka), using the *f2* criterion and by assuming a Near East, Iran and Pakistan source for migrations into Arabian Peninsula. The haplogroup affiliations of the founders are indicated in the bottom.
(TIF)

S26 Fig. Schematic tree of haplogroups L4 and L6. Ages (in ka) indicated are maximum likelihood estimates obtained with the whole-mtDNA genome.
(TIF)

S27 Fig. Frequency maps based on HVS-I data for haplogroups L4 (A) and L6 (B).
(TIF)

S28 Fig. Phylogenetic tree of haplogroup L4a. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S29 Fig. Phylogenetic tree of haplogroup L4b. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S30 Fig. Phylogenetic tree of haplogroup L6. Labels on the branches represent nucleotide positions of transitions, and transversions when followed by a suffix "A," "G," "C," or "T"; reversions by "!"; green indicates synonymous, brown non-synonymous, yellow other coding region, and black control region substitutions. Individual identification is indicated as well as the geographic origin when known (geographic regions are grouped by colour code according to the key). Near the nodes, the TMRCA is indicated (mean and 95% confidence interval) for ρ based on whole-mtDNA sequences (in black), ρ based on synonymous diversity (in green) and for maximum likelihood (in blue).
(TIF)

S31 Fig. Bayesian Skyline Plot (BSP), indicating the median of the hypothetical effective population size through time based on data from haplogroup L4 (A) and haplogroups L4 and L6 (B), assuming a generation time of 25 years.
(TIF)

S32 Fig. Probabilistic proportion of founder clusters considering three migration periods (1.0, 2.5 and 13.0 ka), using the *f1* criterion and assuming an African source for migrations into Arabian Peninsula plus the Near East and Iran. The haplogroup affiliations of the founders are indicated in the bottom.
(TIF)

S33 Fig. Probabilistic proportion of founder clusters considering three migration periods (1.0, 2.5 and 13.0 ka), using the *f2* criterion and assuming an African source for migrations into Arabian Peninsula plus Near East and Iran. The haplogroup affiliations of the founders

are indicated in the bottom.
(TIF)

S34 Fig. Probabilistic proportion of founder clusters considering three migration periods (2.0, 10.0 and 16.0 ka), using the *f1* criterion and assuming Arabian Peninsula plus Near East and Iran migrations into eastern Africa. The haplogroup affiliations of the founders are indicated in the bottom.
(TIF)

S35 Fig. Probabilistic proportion of founder clusters considering three migration periods (2.0, 10.0 and 16.0 ka), using the *f2* criterion and assuming Arabian Peninsula plus Near East and Iran migrations into eastern Africa. The haplogroup affiliations of the founders are indicated in the bottom.
(TIF)

S36 Fig. Probabilistic proportion of founder clusters considering three migration periods (2.0, 10.0 and 16.0 ka), using *f1* criterion and assuming Arabian Peninsula plus Near East and Iran migrations into North Africa. The haplogroup affiliations of the founders are indicated in the bottom.
(TIF)

S37 Fig. Probabilistic proportion of founder clusters considering three migration periods (2.0, 10.0 and 16.0 ka), using the *f2* criterion and assuming Arabian Peninsula plus Near East and Iran migrations into North Africa. The haplogroup affiliations of the founders are indicated in the bottom.
(TIF)

S38 Fig. Population structure inferred by ADMIXTURE analysis. Each individual is represented by a vertical (100%) stacked column of genetic components proportions shown in colour for K = 3, 4 and 5.
(TIF)

S1 Table. Haplotypes for whole-mtDNA sequences that were fully characterised in this study and the corresponding geographic region.
(DOCX)

S2 Table. Published whole-mtDNA sequences used in all phylogenetic tree with the corresponding origin and subclade affiliation.
(DOCX)

S3 Table. Diversity values of ρ and π used for the interpolation maps of the haplogroups J, T and L4.
(DOCX)

S4 Table. Frequency values used in the reconstruction of the interpolation maps for the haplogroups J, T, J1d1a and J2a2b.
(DOCX)

S5 Table. Frequency values used in the reconstruction of the interpolation maps for the haplogroups L4 and L6.
(DOCX)

S6 Table. Founder lineages identified when using *f1* criterion from the Near East, Iran and Pakistan to Arabian Peninsula.
(DOCX)

S7 Table. Founder lineages identified when using *f2* criterion from the Near East, Iranand Pakistan to Arabian Peninsula.
(DOCX)

S8 Table. Founder lineages identified when using a *f1* criterion from Near East, Iran and Pakistan to Arabian Peninsula, based on whole-mtDNA JT sequences.
(DOCX)

S9 Table. Founder lineages identified when using *f1* criterion from Africa to Arabian Peninsula, Near East and Iran.
(DOCX)

S10 Table. Founder lineages identified when using *f2* criterion from Africa to Arabian Peninsula, Near East and Iran.
(DOCX)

S11 Table. Founder lineages identified when using *f1* criterion from Arabian Peninsula, Near East and Iran to North Africa and to eastern Africa separately.
(DOCX)

S12 Table. Founder lineages identified when using *f2* criterion from Arabian Peninsula, Near East and Iran to North Africa and to eastern Africa separately.
(DOCX)

S13 Table. Samples used for genome-wide autosomal analysis.
(DOCX)

S14 Table. Peaks of rate of population size change through time as obtained from the BSPs and periods of time where the rate of population size increase was of at least one individual per 100 individuals in a period of 100 years. Increment ratio corresponds to the number of times the effective population size increase during this period.
(DOCX)

S15 Table. Ages for the oldest founders in the migration from Africa into the Arabian Peninsula, Near East and Iran. This is a sub-set of S9 Table.
(DOCX)

S1 Text. Phylogeographic analyses and supplemental information on founder analyses. Includes 15 tables.
(DOCX)

Author Contributions

Conceived and designed the experiments: MBR LP. Performed the experiments: VF JBP TR AM ZF. Analyzed the data: VF PT BC PS. Contributed reagents/materials/analysis tools: FA VC MBR LP. Wrote the paper: VF MBR LP.

References

1. Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, et al. (2011) The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet 7: e1001373. doi: 10.1371/journal.pgen.1001373 PMID: 21533020

2. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587. PMID: 12930761
3. Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, et al. (2014) A genetic atlas of human admixture history. Science 343: 747–751. doi: 10.1126/science.1243518 PMID: 24531965
4. Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M (2011) Dating the age of admixture via wavelet transform analysis of genome-wide data. Genome Biol 12: R19. doi: 10.1186/gb-2011-12-2-r19 PMID: 21352535
5. Pool JE, Nielsen R (2009) Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181: 711–719. doi: 10.1534/genetics.108.098095 PMID: 19087958
6. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, et al. (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci U S A 107: 786–791. doi: 10.1073/pnas.0909559107 PMID: 20080753
7. Hodgson JA, Mulligan CJ, Al-Meeri A, Raaum RL (2014) Early Back-to-Africa Migration into the Horn of Africa. PLoS Genet 10: e1004393. doi: 10.1371/journal.pgen.1004393 PMID: 24921250
8. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, et al. (2012) Ancient admixture in human history. Genetics 192: 1065–1093. doi: 10.1534/genetics.112.145037 PMID: 22960212
9. Soares P, Ermini L, Thomson N, Mormina M, Rito T, et al. (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84: 740–759. doi: 10.1016/j.ajhg.2009.05.001 PMID: 19500773
10. Fu Q, Mittnik A, Johnson PL, Bos K, Lari M, et al. (2013) A revised timescale for human evolution based on ancient mitochondrial genomes. Curr Biol 23: 553–559. doi: 10.1016/j.cub.2013.02.044 PMID: 23523248
11. Busby GB, Brisighelli F, Sanchez-Diz P, Ramos-Luis E, Martinez-Cadenas C, et al. (2012) The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. Proc Biol Sci 279: 884–892. doi: 10.1098/rspb.2011.1044 PMID: 21865258
12. Wei W, Ayub Q, Chen Y, McCarthy S, Hou Y, et al. (2013) A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res 23: 388–395. doi: 10.1101/gr.143198.112 PMID: 23038768
13. Poznik GD, Henn BM, Yee MC, Sliwerska E, Euskirchen GM, et al. (2013) Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science 341: 562–565. doi: 10.1126/science.1237619 PMID: 23908239
14. Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, et al. (2013) Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 341: 565–569. doi: 10.1126/science.1237947 PMID: 23908240
15. Richards M, Macaulay V, Hickey E, Vega E, Sykes B, et al. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67: 1251–1276. PMID: 11032788
16. Soares P, Alshamali F, Pereira JB, Fernandes V, Silva NM, et al. (2012) The expansion of mtDNA haplogroup L3 within and out of Africa. Mol Biol Evol 29: 915–927. doi: 10.1093/molbev/msr245 PMID: 22096215
17. Soares P, Trejaut JA, Loo JH, Hill C, Mormina M, et al. (2008) Climate change and postglacial human dispersals in southeast Asia. Mol Biol Evol 25: 1209–1218. doi: 10.1093/molbev/msn068 PMID: 18359946
18. Forster P, Harding R, Torroni A, Bandelt HJ (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59: 935–945. PMID: 8808611
19. Scally A, Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet 13: 745–753. doi: 10.1038/nrg3295 PMID: 22965354
20. Mellars P, Gori KC, Carr M, Soares PA, Richards MB (2013) Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. Proc Natl Acad Sci U S A 110: 10699–10704. doi: 10.1073/pnas.1306043110 PMID: 23754394
21. Rito T, Richards MB, Fernandes V, Alshamali F, Cerny V, et al. (2013) The first modern human dispersals across Africa. PLoS One 8: e80031. doi: 10.1371/journal.pone.0080031 PMID: 24236171
22. Costa MD, Pereira JB, Pala M, Fernandes V, Olivieri A, et al. (2013) A substantial prehistoric European ancestry amongst Ashkenazi maternal lineages. Nat Commun 4: 2543. doi: 10.1038/ncomms3543 PMID: 24104924
23. Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, et al. (2010) The genome-wide structure of the Jewish people. Nature 466: 238–242. doi: 10.1038/nature09103 PMID: 20531471
24. Fernandes V, Alshamali F, Alves M, Costa MD, Pereira JB, et al. (2012) The Arabian Cradle: Mitochondrial Relicts of the First Steps along the Southern Route out of Africa. Am J Hum Genet 90: 347–355. doi: 10.1016/j.ajhg.2011.12.010 PMID: 22284828

25. Petraglia MD, Alsharekh A (2003) The Middle Palaeolithic of Arabia: Implications for modern human origins, behaviour and dispersals Antiquity 77: 671–684

26. Parker AG (2009) Pleistocene Climate Change in Arabia: Developing a Framework for Hominin Dispersal over the Last 350 ka. In: Petraglia MD, Rose J, editors. The Evolution of Human Populations in Arabia: Paleoenvironments, Prehistory and Genetics. The Netherlands: Springer. pp. 39–50.

27. Rose J, Petraglia MD (2009) Tracking the Origin and Evolution of Human Populations in Arabia. In: Petraglia MD, Rose J, editors. The Evolution of Human Populations in Arabia: Paleoenvironments, Prehistory and Genetics. The Netherlands: Springer. pp. 1–14.

28. Drechsler P (2009) The dispersal of the Neolithic over the Arabian Peninsula. Archaeopress, Oxford: British Archaeological Reports International Series S1969.

29. Groucutt HS, Petraglia MD (2012) The prehistory of the Arabian peninsula: deserts, dispersals, and demography. Evol Anthropol 21: 113–125. doi: 10.1002/evan.21308 PMID: 22718479

30. Uerpmann H-P, Potts DT, Uerpmann M (2009) Holocene (Re-) Occupation of Eastern Arabia. In: Petraglia MD, Rose J, editors. The Evolution of Human Populations in Arabia: Paleoenvironments, Prehistory and Genetics. The Netherlands: Springer. pp. 205–214.

31. Fedele FG (2009) Early Holocene in the Highlands: Data on the Peopling of the Eastern Yemen Plateau, with a Note on the Pleistocene Evidence. In: Petraglia MD, Rose J, editors. The Evolution of Human Populations in Arabia: Paleoenvironments, Prehistory and Genetics. The Netherlands: Springer. pp. 215–236.

32. Cerny V, Mulligan CJ, Fernandes V, Silva NM, Alshamali F, et al. (2011) Internal diversification of mitochondrial haplogroup R0a reveals post-last glacial maximum demographic expansions in South Arabia. Mol Biol Evol 28: 71–78. doi: 10.1093/molbev/msq178 PMID: 20643865

33. Boivin N, Blench R, Fuller DQ (2009) Archaeological, Linguistic and Historical Sources on Ancient Seafaring: A Multidisciplinary Approach to the Study of Early Maritime Contact and Exchange in the Arabian peninsula. In: Petraglia MD, Rose J, editors. The Evolution of Human Populations in Arabia: Paleoenvironments, Prehistory and Genetics. The Netherlands: Springer. pp. 251–278.

34. Mitchell P (2005) African connections: archaeological perspectives on Africa and the wider world. Walnut Creek: Altamira Press. pp. 328.

35. Ray PH (2003) The Archaeology of seafaring in ancient South Asia. Cambridge: Cambridge University Press. pp350.

36. Kitchen A, Ehret C, Assefa S, Mulligan CJ (2009) Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. Proc Biol Sci 276: 2703–2710. doi: 10.1098/rspb.2009.0408 PMID: 19403539

37. Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, et al. (2004) Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75: 752–770. PMID: 15457403

38. Lovejoy PE (1983) Transformations inSlavery—A history of slavery in Africa. Third Edition ed. New York: Cambridge University Press. pp. 200.

39. Segal R (2002) Islam's Black Slaves—the other black diaspora. London: Atlantic Books. pp 288.

40. Richards M, Rengo C, Cruciani F, Gratrix F, Wilson JF, et al. (2003) Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations. Am J Hum Genet 72: 1058–1064. PMID: 12629598

41. Freitag U, Clarence-Smith WG (1997) Hadhrami traders, scholars, and statesmen in the Indian Ocean. Leiden; New York: Brill. pp 392.

42. Alshamali F, Brandstatter A, Zimmermann B, Parson W (2008) Mitochondrial DNA control region variation in Dubai, United Arab Emirates. Forensic Sci Int Genet 2: e9–10. doi: 10.1016/j.fsigen.2007.11.001 PMID: 19083802

43. Abu-Amero KK, Gonzalez AM, Larruga JM, Bosley TM, Cabrera VM (2007) Eurasian and African mitochondrial DNA influences in the Saudi Arabian population. BMC evolutionary biology 7: 32. PMID: 17331239

44. Abu-Amero KK, Larruga JM, Cabrera VM, Gonzalez AM (2008) Mitochondrial DNA structure in the Arabian Peninsula. BMC evolutionary biology 8: 45. doi: 10.1186/1471-2148-8-45 PMID: 18269758

45. Cerny V, Mulligan CJ, Ridl J, Zaloudkova M, Edens CM, et al. (2008) Regional differences in the distribution of the sub-Saharan, West Eurasian, and South Asian mtDNA lineages in Yemen. Am J Phys Anthropol 136: 128–137. doi: 10.1002/ajpa.20784 PMID: 18257024

46. Cerny V, Pereira L, Kujanova M, Vasikova A, Hajek M, et al. (2009) Out of Arabia-the settlement of island Soqotra as revealed by mitochondrial and Y chromosome genetic diversity. Am J Phys Anthropol 138: 439–447. doi: 10.1002/ajpa.20960 PMID: 19012329

47. Al-Abri A, Podgorna E, Rose JI, Pereira L, Mulligan CJ, et al. (2012) Pleistocene-Holocene boundary in Southern Arabia from the perspective of human mtDNA variation. Am J Phys Anthropol 149: 291–298. doi: 10.1002/ajpa.22131 PMID: 22927010

48. Musilova E, Fernandes V, Silva NM, Soares P, Alshamali F, et al. (2011) Population history of the Red Sea—genetic exchanges between the Arabian Peninsula and East Africa signaled in the mitochondrial DNA HV1 haplogroup. Am J Phys Anthropol 145: 592–598. doi: 10.1002/ajpa.21522 PMID: 21660931

49. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100–1104. doi: 10.1126/science.1153717 PMID: 18292342

50. The, Genomes, Project, Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073. doi: 10.1038/nature09534 PMID: 20981092

51. Arredi B, Poloni ES, Paracchini S, Zerjal T, Fathallah DM, et al. (2004) A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. Am J Hum Genet 75: 338–345. PMID: 15202071

52. Kujanova M, Pereira L, Fernandes V, Pereira JB, Cerny V (2009) Near eastern neolithic genetic input in a small oasis of the Egyptian Western Desert. Am J Phys Anthropol 140: 336–346. doi: 10.1002/ajpa.21078 PMID: 19425100

53. Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, et al. (2006) The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. Science 314: 1767–1770. PMID: 17170302

54. Pereira L, Silva NM, Franco-Duarte R, Fernandes V, Pereira JB, et al. (2010) Population expansion in the North African late Pleistocene signalled by mitochondrial DNA haplogroup U6. BMC Evol Biol 10: 390. doi: 10.1186/1471-2148-10-390 PMID: 21176127

55. Pennarun E, Kivisild T, Metspalu E, Metspalu M, Reisberg T, et al. (2012) Divorcing the Late Upper Palaeolithic demographic histories of mtDNA haplogroups M1 and U6 in Africa. BMC Evol Biol 12: 234. doi: 10.1186/1471-2148-12-234 PMID: 23206491

56. Harich N, Costa MD, Fernandes V, Kandil M, Pereira JB, et al. (2010) The trans-Saharan slave trade—clues from interpolation analyses and high-resolution characterization of mitochondrial DNA lineages. BMC Evol Biol 10: 138. doi: 10.1186/1471-2148-10-138 PMID: 20459715

57. Cerezo M, Achilli A, Olivieri A, Perego UA, Gomez-Carballa A, et al. (2012) Reconstructing ancient mitochondrial DNA links between Africa and Europe. Genome Res 22: 821–826. doi: 10.1101/gr.134452.111 PMID: 22454235

58. Cerny V, Pereira L, Musilova E, Kujanova M, Vasikova A, et al. (2011) Genetic structure of pastoral and farmer populations in the African Sahel. Mol Biol Evol. 28:2491–500. doi: 10.1093/molbev/msr067 PMID: 21436121

59. Pereira L, Prata MJ, Amorim A (2000) Diversity of mtDNA lineages in Portugal: not a genetic edge of European variation. Ann Hum Genet 64: 491–506. PMID: 11281213

60. van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30: E386–394. doi: 10.1002/humu.20921 PMID: 18853457

61. Kloss-Brandstatter A, Pacher D, Schonherr S, Weissensteiner H, Binna R, et al. (2011) HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat 32: 25–32. doi: 10.1002/humu.21382 PMID: 20960467

62. Pereira L, Goncalves J, Franco-Duarte R, Silva J, Rocha T, et al. (2007) No evidence for an mtDNA role in sperm motility: data from complete sequencing of asthenozoospermic males. Mol Biol Evol 24: 868–874. PMID: 17218641

63. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, et al. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23: 147. PMID: 10508508

64. Bandelt HJ, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141: 743–753. PMID: 8647407

65. Pereira L, Freitas F, Fernandes V, Pereira JB, Costa MD, et al. (2009) The diversity present in 5140 human mitochondrial genomes. Am J Hum Genet 84: 628–640. doi: 10.1016/j.ajhg.2009.04.013 PMID: 19426953

66. Soares P, Ermini L, Thomson N, Mormina M, Rito T, et al. (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84: 740–759. doi: 10.1016/j.ajhg.2009.05.001 PMID: 19500773

67. Saillard J, Forster P, Lynnerup N, Bandelt HJ, Norby S (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67: 718–726. PMID: 10924403

68. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556. PMID: 9367129

69. Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22: 1185–1192. PMID: 15703244

70. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7: 214. PMID: 17996036

71. Atkinson QD, Gray RD, Drummond AJ (2009) Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa. Proc Biol Sci 276: 367–373. doi: 10.1098/rspb.2008.0785 PMID: 18826938

72. Alves M, Alves J, Camacho R, Soares P, Pereira L. From Networks to Trees. In: Springer, editor; 2012; Salamanca-Spain. pp. 129–136.

73. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575. PMID: 17701901

74. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19: 1655–1664. doi: 10.1101/gr.094052.109 PMID: 19648217

75. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190. PMID: 17194218

4 Final Discussion

Slavery has been an unfortunate part of human history. Virtually all great empires practiced slavery, as their economy often depended on slave workforce. Although the slavery was practiced differently between cultures, there are several features that are common to basically all known episodes of slavery and slave trading. For our work, two particularly features are important: large scale forced migrations and subsequent sex-biased admixture. If the episode of slave trading occurs between two genetically divergent populations, later admixture produces a particular genetic structure in the descendant population. In the cases of Trans-Atlantic Slave Trade and Arab Slave Trade, the admixture took place between populations with considerably distinct gene pools and therefore these African slave trading systems offer unique opportunities to study human history from the perspective of genetics. In addition, mixture of African and Eurasian genetic backgrounds inevitably brings health related implications: the African genetic background confers better resistance to certain infectious diseases (Karlsson et al. 2014), but it also brings susceptibility to blood disorders, diabetes, kidney failure and other (Palmer et al. 2012, Franceschini et al. 2013).

Population structure

We began this work by characterizing the population structure and evolutionary adaptations in the Sahel Belt, Africa's most important migration corridor and genetic "melting pot", directly connected to Trans-Atlantic Slave Trade system in the West, and Arab Slave Trade in the East. We have observed three main clusters of genetic diversity in the Sahel Belt: Eastern, heavily influenced by Eurasian admixture, Atlantic West and West/Central African components. Interestingly, we distinguished the same clusters of African genetic diversity in the Cuban population: the West/Central African cluster was dominant (66-70% of the total African ancestry), followed by the Atlantic West African (20-24%) and with minor traces of East African ancestry (10%). These findings are in line with historical records: slaves brought to Cuba were brought mainly from ports in West and Central Africa, and a minor portion from Senegambian coast (Eltis & Richardson 2010).

The substructure in the West African genetic component between Atlantic West Africa (Senegal and Gambia) and West/Central Africa (Gulf of Guinea) has been described only recently when high-density genotyping platforms were employed (Gurdasani et al. 2015). Our data, based on 2.5 million SNP, confirmed this substructure. Also, all investigated West African populations derive their ancestry from both these sources, without evidence of admixture with other population backgrounds.

Particularly interesting is Eurasian admixture in Sub-Saharan Africa. All investigated

populations from Eastern and Central Sahel contain portions of Eurasian genetic ancestry. We distinguished two different components of Eurasian admixture in Sahel populations: one widely present in Eastern Sahel and apparently related to Near Eastern/Arabian Peninsula genetic background; and another related to autochthonous North Africans, present in the nomadic populations from Central and Western Sahel. These two Eurasian ancestries were brought to Africa during episodes of Back-to-Africa migrations during favourable climatic conditions in the Near East (Rose & Petraglia 2009). While in the East Africa the main migration took place in Late Glacial period around 16 ka, in North Africa the main Back-to-Africa migration occurred during the Neolithic (our work).

Time of admixture can be estimated through two different strategies: i) methods that rely on phylogenetic analysis of uniparental markers accompanied by dating by molecular clock and ii) methods measuring decay of LD and attempting to fit the LD decay with regression curve in order to estimate the time (Moorjani et al. 2011). Each of these methods is better suited for particular situations: LD based methods perform well in situations of relatively recent admixture before the LD decays completely; methods based on analysis of uniparental markers perform arguably better in detecting very old admixtures and multiple waves of migration.

We employed both approaches to estimate the dates of Eurasian admixture in Africa, and of African admixture in the Arabian Peninsula. Results of ADMIXTURE analysis suggested that the Eurasian ancestry in Daza, Kanembu and Fulani most probably originated from North Africa, which is corroborated also by evidence from mitochondrial data (Cerný et al. 2006; Podgorná et al. 2013). Ages of admixture for these populations estimated by ALDER yielded considerably more recent estimates for Fulani (17.40 ± 4.63 generations) than for Daza+Kanembu (39.85 ± 4.41), suggesting that these admixture processes resulted from different demographic events. This fact is corroborated by mtDNA data: Eurasian mitochondrial haplogroups found in Fulani comprise J1b, U5, H and V (Cerny et al. 2006), while Eurasian haplogroups in Daza are represented by M1 and U6 (Podgorna et al. 2013).

History of Eurasian admixture in East Africa seems to be even more complex. The archaeological record provides evidence for contact between the Horn of Africa and Near East, beginning 3,000 year ago (Phillipson 2010). A similar time of admixture was obtained by LD based methods, and was assumed as the time of Eurasian admixture in Africa (Pagani et al. 2012, Pickrell et al. 2014). However, Hodgson et al. (2014) pointed out that the LD based methods are strongly biased towards most recent admixture events and therefore perform poorly in scenarios with several successive migrations. The latter study attested for earlier

Back-to-Africa migration into the Horn of Africa, probably in pre-agricultural era. Our results of founder analysis confirm that the Late Glacial period was the time of major Back-to-Africa migration, although there were at least other two back migrations into Africa, one around the early Neolithic, and second very recently within last thousand years.

The difference between Eastern and Western Sahel observed on population structure is reflected also in the patterns of positive selection. In general, West African populations exhibit extensive sharing of selection signals, while on Eastern Sahel populations these signals appear to be more diverse. It is important to note, that except Fulani, our West African populations are sedentary, while in Central and Eastern Sahel we have sampled both nomadic or semi-nomadic pastoralists, as well as sedentary populations. Pathway analysis of iHS and XP-EHH results indicated that the West African populations have been targeted by selection on heart and oxytocin pathways, while Eastern Sahel populations display selection on lipid metabolism pathways, namely on glycerolipid and glycerophospholipid pathways.

Differences between Eastern and Western Sahel patterns of selection are also apparent in results of XP-EHH test, where West African populations reached high values among *SPINT2* and *CATSPERG* region, associated with diarrhea (Heinz-Erian et al. 2008) and male fertility (Wang et al. 2004), respectively. In Eastern Sahel, top scores in XP-EHH vs West were for *DGAT2* and *DGKI* genes, both playing key roles in glycerolipid and glycerophospholipid pathways (Yen et al. 2008).

Although a large proportion of reported signals was geographically clustered within Eastern or Western Sahel, we have observed also a number of signals universally presented by all investigated populations. A clear example is *DARC* gene associated with resistance to *P. vivax* malaria, through absence of Duffy antigen on erythrocyte surface (Nickel et al. 1999). Another example of across Sahel selected gene is *PIGG*, which is involved in the formation of GPI-anchor, a cell membrane structure allowing the attachment of proteins to the membrane (Stokes et al. 2014).

Local ancestry inference as a tool to identify evolutionary adaptation

Greater genetic diversity in populations mixed from different ancestral backgrounds gives them an advantage when responding to selective pressure forces. If an advantageous allele is present in one of the ancestral backgrounds, this particular background would become favoured by evolution in the genomic region where the advantageous allele lies. This results in a genomic signature of excess of local ancestry in the advantageous region. Compared to classical genome-wide association chi-square test, admixture mapping has considerably

higher power due to the lowered testing burden, since the ancestral blocks typically span over tens to thousands of SNPs (Montana & Hoggart 2007).

We have successfully applied the approach of admixture mapping to the populations in Eastern Sahel and Cuba, where Eurasian and African gene pools have been mixing in the course of history, for a considerable longer time in the first case. Because African and Eurasian ancestries are considerably divergent, the identification of the ancestral blocks is typically performed with high accuracy. We have found excess of African ancestry in genomic regions capable of influencing contraction and progress of infectious diseases. Because of the strong pathogen-driven selective pressure in Africa, the African genetic background accumulated bulk of genetic variants that confer resistance to certain infectious diseases. An excellent example from our data is represented by *DARC* region enriched for African ancestry in Sudanese Arabs and Nubians, originally Eurasian populations (modern Sudanese Arabs and Nubians are approximately 50% Eurasian), residing in Africa only for several hundreds of years. Assuming that Arabs and Nubians had only little, if any, innate resistance to malaria before they arrived to the Sahel Belt, individuals who inherited the African allele of *DARC* gene had strong advantage. The selection on *DARC* has been already described (Sabeti et al. 2007), but we have demonstrated its importance in the context of East African adaptative admixture.

Similarly, we have observed excess of African ancestry in *RXRA-COL5A1* region in the Cuban Dengue cohort, when comparing asymptomatic/controls with haemorrhagic patients. As we will discuss later, this region (particularly *RXRA*) seems to play a role in the regulation of lipid metabolism pathways.

On the contrary, we have also observed excess in Eurasian local ancestry in admixed African populations. In Eastern Sahel populations, the *RAB3GAP1*/*LCT*/*MCM6* region on chromosome 2 is significantly enriched for Eurasian ancestry, suggesting that Eurasian genetic variants in this region are particularly advantageous compared to African variants. This region spans across 830 kb in 2q21.3 and harbours several genes known to be under positive selection. The region between *LCT* and *MCM6* contains variants associated with lactase persistence, one of the strongest selection sweeps discovered so far, particularly in northern European populations (Sabeti et al. 2007; Karlsson et al. 2014). However, our results do not support *LCT* as a driver of selection, because none of our Eastern Sahel samples had the European allele for lactase persistence. We propose that the selection in this region is rather being driven by lipid metabolism pathways. This is consistent with results of positive selection on glycerolipid and glycerophospholipid metabolism pathways, as well as strong iHS values for *RAB3GAP1* gene.

Additionally, this region contains microRNA-128 involved in cholesterol and lipid metabolism, as well as insulin sensitivity (Naar & Najafi-Shoushtari 2013). To sum up, several lines of evidence suggest that lipid metabolism was targeted in Eastern Sahel populations. As proposed previously (Wagh et al. 2012), this could be related to diet, although, as we also show below in the case study in Cuba, lipid metabolism plays a crucial role in host response to the infection (Joseph et al. 2003; Ma et al. 2014).

Particularly interesting is the local ancestry pattern in Fulani. We detected excess of Eurasian ancestry at chromosome 12 in the region of taste receptors type 2, responsible for perception of bitter taste. In our dataset this signal is restricted only to Fulani population. Previous studies have shown that the members of *TAS2R* family are under positive selection in Eurasian and African populations, although the patterns of selection and the particular selected genes are different and located in another chromosome (Wang et al. 2004). It has been previously assumed that the selection on members like *TAS2R16*, which code for sensing of bitter anti-inflammatory compound salicin, could be driven by the ability to digest medication (Campbell et al. 2014). Nevertheless, in our work we report the region with *TAS2R* members responsible for sensing natural alkaloids (Ledda et al. 2014). It is worth to note, that the cultural tradition of Fulani embraces rituals that involve ingurgitation of a bitter beverage containing seeds of the plant *Datura metel.* As the initiation ritual of boys (Sharo) is painful, participants use this beverage for its narcotic effect provided by alkaloids scopolamine or hyoscine, which depress the central nervous system (Adeola 2014). Because any sign of pain or tearing will disadvantage young Fulani men in competition for women, we propose that the ability to digest this bitter beverage might be actually targeted by sexual selection.

Role of African ancestry background in lipid metabolism pathways

The work performed in this thesis underlines the importance of lipid metabolism in the evolutionary adaptation in African and in African descendant populations. Although our motivation and approach was partly different in each of the studies, in both studies the lipid metabolism emerged as a key target of evolution in those populations. Despite the fact that the genes and metabolic pathways we identified in the Cuban case and control groups are different from those seen in the Sahel populations, there are several points connecting them together, which we will discuss below.

In East Africa, we observed strong signals of positive selection on genes involved in glycerolipid and glycerophospholipid metabolism, and on *RAB3GAP* gene, which is involved in the regulation of blood lipids (Teslovich et al. 2010). However, both of these pathways are also

involved in a vast number of body functions and it is difficult to pinpoint the exact force that has been driving the selection, especially when it is plausible that various selective forces acted upon some of the identified candidate regions. We have proposed that in East Africa, glycerolipid, glycerophospholipid and cholesterol metabolism pathways could be under selective pressure because of the specific diet rich in cholesterol, which is characteristic of some populations, e.g. Maasai (Biss et al. 1971). Similar signals were observed also by Wagh et al. (2012), arriving to the conclusion that the selection on the lipid pathway was probably driven by the extreme diet rich in meat. Nevertheless, these pathways are possibly targeted by several evolutionary forces, since the glycerolipid and glycerophospholipid metabolism pathways play a role also in the formation of cell membrane (Farooqui et al. 2000), cell differentiation and cell survival (Prentki and Madiraju 2008). For example, *CD36* is one of the genes implicated in glycerolipid metabolism, but has been known to confer resistance to malaria (McGilvray et al. 2000). Recently, Durán et al. (2015) demonstrated, that lipid levels are directly linked to outcome of dengue infection, what is particularly relevant for our case study in Cuba.

In the context of African and Eurasian admixture, it is important to note that the lipid profiles of Europeans and Africans are different: Africans have lower cholesterol levels and higher levels of low density lipids (LDL) (Goedecke et al. 2010). Curiously, the genes that we identified in the association study and admixture mapping in Cuba (*RXRA* and *OSBPL10)* have different genetic variants in Africans and Europeans and also the expression levels of *OSBPL10* are different, suggesting that this gene supposedly contribute to differentiated lipid profiles between Africans and Europeans.

The role of lipids in dengue infection is possibly related to the virus replication and maturation, which take place on endoplasmatic reticulum and Golgi complex. In these cellular structures, the oxysterol-binding proteins (OSBPs) play a key role in cholesterol homeostasis (Du et al. 2015). The second part of the cholesterol homeostasis is mediated through LXR/RXR activation pathway. *RXRA* regulates lipid metabolism in macrophages, the main target cells of dengue virus and LXR/RXRA heterodimers inhibit *NF-κB* gene which acts upon inflammatory mediators (Joseph et al. 2003). On the other hand, LXR/RXR dimers are negatively controlled by *IRF3* when pathogens enter the cell through the toll-like receptors (TLRs), allowing proper action of *NF-κB*. During dengue infection, the expression of *RXRA* is downregulated, so the expression of type I interferon (*IFN*) could reach optimal levels, and after the infection the levels of *RXRA* return to normal (Ma et al. 2014). The fact that the downregulation of *RXRA* improves antiviral response is in line with the measured levels of mRNA expression in dengue patients

along the infection process: expression of *RXRA* was significantly lower during the first phases of infection and higher in healthy and convalescent patients, both in Cuba and in Thailand.

Our hypothesis about the central role of lipid metabolism in dengue infection provides a synthetic explanation of previously discovered associations with dengue. The protein product of the *VDR* gene, which has been associated with protective effect in the Vietnamese population (Loke et al. 2002), forms heterodimers with *RXRA* and suppresses function of *NF-κB* (D'Ambrosio et al. 1998). In addition, the *PLCE1* gene, associated with protection against DSS in Vietnamese children (Khor et al. 2011), is implicated in the PPARA/RXRA activation pathway while affecting lipid metabolism (Motojima et al. 1998). We propose, that the protective variants in *VDR* and *PLCE1* arose in Asian populations, while in African populations the protection mediated by *RXRA* and *OSBPL10* were selected. The connecting point is the RXRA-various dimers based-pathways, which play a role in lipid signaling and chemokine production.

In fact, a link can also be made between the African enriched genomic regions found in Cuban asymptomatic patients and in Eastern Sahel, through *RXRA* in PPAR activation pathways. There are three classes of PPAR receptors: PPARA are present in skeletal muscles and liver, PPARB in skeletal muscles and adipocytes and PPARG in adipocytes. All three of them use 9-cis-retinoic acid for activation of the heterodimers PPAR-RXR that target expression of numerous genes implicated in lipid metabolism, including the glycerophospholipid metabolism (**Figure 16**).

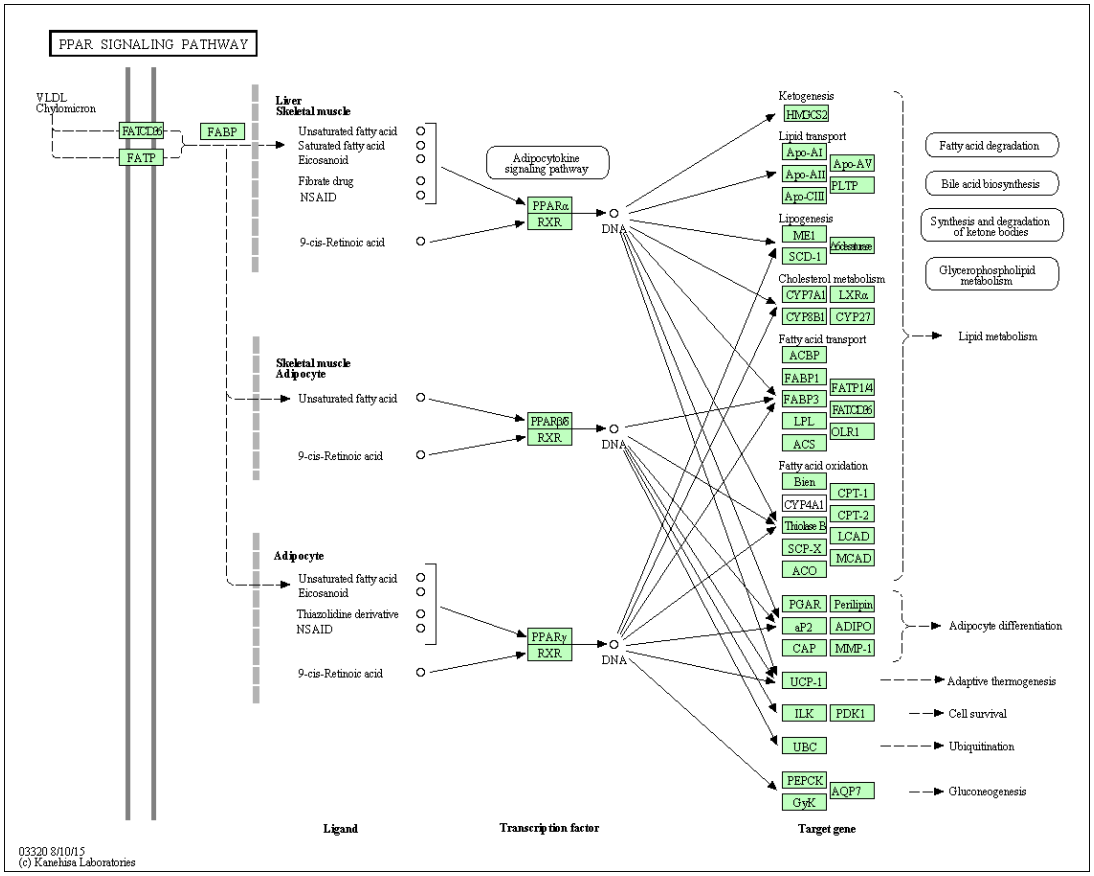

Figure 16 - PPAR signalling pathway. Adapted from KEGG: Kyoto Encyclopedia of Genes and Genomes (Ogata et al. 1999).

5 Concluding remarks

The work presented in this thesis has several important implications for research of human history, evolution and adaptation to infectious diseases.

1. **The Trans-Atlantic Slave Trade and the Arab Slave Trade redefined structure of populations in non-African regions.** We confirmed that various African populations contributed to the gene pool of American populations and, in the Cuban case study, we elucidated the role played by the African ancestry on the progress of dengue infection. We also illuminated the patterns of African admixture in Arabian Peninsula, where health-related issues of African ancestry arise from various types of blood disorders of African origin.

2. **Lipid metabolism has been targeted by positive selection in different African and African-descendant populations.** We identified signals of positive selection on numerous genes involved in lipid metabolism in East African populations, supposedly related to a diet rich in meat, blood and milk. In addition, we found an association with outcome of dengue infection on a different set of genes related to lipid metabolism in the Cuban admixed population.

3. **Identified natural strategies of adaptation to complex diseases open new ways for pharmacological research.** Methods for targeting microRNA in *RAB3GAP-LCT* region have been already patented with prospective of development new treatments of insulin resistance and lipid metabolism regulation. Our results may foster research also on the glycerolipid metabolism pathway, and dengue research community may start targeting RXR pathway in search for better treatment methods.

6 References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. Nature 526:68–74.

Abu-Amero KK, Hellani A, González AM, Larruga JM, Cabrera VM, Underhill PA. 2009. Saudi Arabian Y-Chromosome diversity and its relationship with nearby regions. BMC Genet. 10:59.

Adeola BS. 2014. Datura metel L.: Analgesic or Hallucinogen? “Sharo” Pespective. Middle-East J. Sci. Res. 21:993–997.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664.

Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L. 2015. Population genomics of Bronze Age Eurasia. Nature 522:167–172.

Alves-Silva J, da Silva Santos M, Guimarães PE, Ferreira AC, Bandelt HJ, Pena SD, Prado VF. 2000. The ancestry of Brazilian mtDNA lineages. Am. J. Hum. Genet. 67:444–461.

Arredi B, Poloni ES, Paracchini S, Zerjal T, Fathallah DM, Makrelouf M, Pascali VL, Novelletto A, Tyler-Smith C. 2004. A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. Am J Hum Genet 75:338–345.

Bachmanov AA, Beauchamp GK. 2007. Taste receptor genes. Annu. Rev. Nutr. 27:389–414.

Barbujani G, Bertorelle G, Chikhi L. 1998. Evidence for Paleolithic and Neolithic gene flow in Europe. Am. J. Hum. Genet. 62:488.

Batista dos Santos SE, Rodrigues JD, Ribeiro-dos-Santos KC, Zago MA. 1999. Differential contribution of indigenous men and women to the formation of an urban population in the Amazon region as revealed by mtDNA and Y-DNA. Am. J. Phys. Anthropol. 109:175–180.

Bedoya G, Montoya P, García J, Soto I, Bourgeois S, Carvajal L, Labuda D, Alvarez V, Ospina J, Hedrick PW, Ruiz-Linares A, Garcı J. 2006. Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. Proc. Natl. Acad. Sci. U. S. A. 103:7234–7239.

Benn-Torres J, Bonilla C, Robbins CM, Waterman L, Moses TY, Hernandez W, Santos ER, Bennett F, Aiken W, Tullock T, Coard K, Hennis a, Wu S, Nemesure B, Leske MC, Freeman V, Carpten J, Kittles R a. 2008. Admixture and population stratification in African Caribbean populations. Ann. Hum. Genet. 72:90–98.

Bernardo S, Hermida R, Desidério M, Silva DA, De Carvalho EF. 2014. MtDNA ancestry of Rio de Janeiro population, Brazil. Mol. Biol. Rep. 41:1945–1950.

Berniell-Lee G, Calafell F, Bosch E, Heyer E, Sica L, Mouguiama-Daouda P, Van Der Veen L, Hombert JM, Quintana-Murci L, Comas D. 2009. Genetic and demographic implications of the Bantu expansion: Insights from human paternal lineages. Mol. Biol. Evol. 26:1581–1589.

Bernstein C, Bernstein H, Payne CM, Garewal H. 2002. DNA repair/pro-apoptotic dual-role proteins in five major DNA repair pathways: fail-safe protection against carcinogenesis. Mutat. Res. 511:145–178.

Biss K, Ho K-J, Mikkelson B, Lewis L, Taylor CB. 1971. Some unique biologic characteristics of the Masai of East Africa. N. Engl. J. Med. 284:694–699.

Bortolini MC, Da Silva WA, De Guerra DC, Remonatto G, Mirandola R, Hutz MH, Weimer TA, Silva MCBO, Zago MA, Salzano FM. 1999. African-derived South American populations:

A history of symmetrical and asymmetrical matings according to sex revealed by bi-and uni-parental genetic markers. Am. J. Hum. Biol. 11:551–563.

Bozzola M, Travaglino P, Marziliano N, Meazza C, Pagani S, Grasso M, Tauber M, Diegoli M, Pilotto A, Disabella E, Tarantino P, Brega A, Arbustini E. 2009. The shortness of Pygmies is associated with severe under-expression of the growth hormone receptor. Mol. Genet. Metab. 98:310–313.

Brucato N, Cassar O, Tonasso L, Tortevoye P, Migot-Nabias F, Plancoulaine S, Guitard E, Larrouy G, Gessain A, Dugoujon J-M. 2010. The imprint of the Slave Trade in an African American population: mitochondrial DNA, Y chromosome and HTLV-1 analysis in the Noir Marron of French Guiana. BMC Evol. Biol. 10:314.

Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo J-M, Wambebe C, Tishkoff S a, Bustamante CD. 2010a. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci. U. S. A. 107:786–791.

Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, Hammer M, Bustamante CD, Ostrer H. 2010b. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proc. Natl. Acad. Sci. U. S. A. 107 Suppl:8954–8961.

Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. 2014. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. Am. J. Hum. Genet. 96:37–53.

Bučková J, Černý V, Novelletto A. 2013. Multiple and differentiated contributions to the male gene pool of pastoral and farmer populations of the African Sahel. Am. J. Phys. Anthropol. 151:10–21.

Campbell MC, Ranciaro A, Zinshteyn D, Rawlings-Goss R, Hirbo J, Thompson S, Woldemeskel D, Froment A, Rucker JB, Omar S a., Bodo JM, Nyambo T, Belay G, Drayna D, Breslin P a S, Tishkoff S a. 2014. Origin and differential selection of allelic variation at TAS2R16 associated with salicin bitter taste sensitivity in Africa. Mol. Biol. Evol. 31:288–302.

Cann RL, Stoneking M, Wilson AC. 1987. Mitochondrial DNA and human evolution. Nature 325:31–36.

Cao A, Galanello R. 2010. Beta-thalassemia. Genet. Med. 12:61–76.

Carvajal-Carmona LG, Soto ID, Pineda N, Ortíz-Barrientos D, Duque C, Ospina-Duque J, McCarthy M, Montoya P, Alvarez VM, Bedoya G, others. 2000. Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. Am. J. Hum. Genet. 67:1287–1295.

Carvalho-Silva DR, Santos FR, Rocha J, Pena SD. 2001. The phylogeography of Brazilian Y-chromosome lineages. Am. J. Hum. Genet. 68:281–286.

Cerný V, Hájek M, Bromová M, Cmejla R, Diallo I, Brdicka R. 2006. MtDNA of Fulani nomads and their genetic relationships to neighboring sedentary populations. Hum. Biol. an Int. Rec. Res. 78:9–27.

Cerný V, Pereira L, Musilová E, Kujanová M, Vašíková A, Blasi P, Garofalo L, Soares P, Diallo I, Brdička R, Novelletto A. 2011. Genetic structure of pastoral and farmer populations in the African Sahel. Mol. Biol. Evol. 28:2491–2500.

Chacón-Duque JC, Adhikari K, Avendaño E, Campo O, Ramirez R, Rojas W, Ruiz-Linares A, Restrepo BN, Bedoya G. 2014. African genetic ancestry is associated with a protective effect on Dengue severity in Colombian populations. Infect. Genet. Evol. 27:89–95.

Chamary J V, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat. Rev. Genet. 7:98–108.

Chinnery PF, Howell N, Andrews RM, Turnbull DM. 1999. Clinical mitochondrial genetics. J. Med. Genet. 36:425–436.

Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT. 2011. Basic statistical analysis in genetic case-control studies. Nat. Protoc. 6:121–133.

Cooke JN, Ng MCY, Palmer ND, An SS, Hester JM, Freedman BI, Langefeld CD, Bowden DW. 2012. Genetic risk assessment of type 2 diabetes-associated polymorphisms in African Americans. Diabetes Care 35:287–292.

Cowie CC, Port FK, Wolfe RA, Savage PJ, Moll PP, Hawthorne VM. 1989. Disparities in incidence of diabetic end-stage renal disease according to race and type of diabetes. N. Engl. J. Med. 321:1074–1079.

Cox G, Sealy J. 1997. Investigating identity and life histories: isotopic analysis and historical documentation of slave skeletons found on the Cape Town foreshore, South Africa. Int. J. Hist. Archaeol. 1:207–224.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Kähäri AK, et al. 2014. Ensembl 2015. Nucleic Acids Res. 43:D662–9.

D'Ambrosio D, Cippitelli M, Cocciolo MG, Mazzeo D, Di Lucia P, Lang R, Sinigaglia F, Panina-Bordignon P. 1998. Inhibition of IL-12 production by 1,25-dihydroxyvitamin D3. Involvement of NF-kappaB downregulation in transcriptional repression of the p40 gene. J. Clin. Invest. 101:252–262.

Dalemans W, Barbry P, Champigny G, Jallat S, Dott K, Dreyer D, Crystal RG, Pavirani A, Lecocq JP, Lazdunski M. 1991. Altered chloride ion channel kinetics associated with the delta F508 cystic fibrosis mutation. Nature 354:526–528.

Darwin C. 1872. The origin of species by means of natural selection; or, The preservation of favored races in the struggle for life. London: Murray. 4th edition. 458pp.

Dawson-Hughes B, Harris SS, Finneran S. 1995. Calcium absorption on high and low calcium intakes in relation to vitamin D receptor genotype. J. Clin. Endocrinol. Metab. 80:3657–3661.

Deshpande DA, Wang WCH, McIlmoyle EL, Robinett KS, Schillinger RM, An SS, Sham JSK, Liggett SB. 2010. Bitter taste receptors on airway smooth muscle bronchodilate by localized calcium signaling and reverse obstruction. Nat. Med. 16:1299–1304.

Diamond J. 1997. Location, location, location: the first farmers. Science. 278:1243–1244.

Diouf SA. 2014. Slavery's exiles: the story of the American Maroons. New York: NYU Press. 403pp.

Dolan MJ, Kulkarni H, Camargo JF, He W, Smith A, Anaya J, Miura T, Hecht FM, Mamtani M, Pereyra F, Marconi V, Mangano A, Sen L, Bologna R, Clark RA, Anderson SA, Delmar J, O'Connell RJ, Lloyd A, et al. 2007. CCL3L1 and CCR5 influence cell-mediated immunity and affect HIV-AIDS pathogenesis via viral entry-independent mechanisms. Nat. Immunol. 8:1324–1336.

Du X, Brown AJ, Yang H. 2015. Novel mechanisms of intracellular cholesterol transport: oxysterol-binding proteins and membrane contact sites. Curr. Opin. Cell Biol. 35:37–42.

Durán A, Carrero R, Parra B, González A, Delgado L, Mosquera J, Valero N. 2015. Association of lipid profile alterations with severe forms of dengue in humans. Arch. Virol.:1–6.

Ehlers J, Gibbard PL. 2003. Extent and chronology of glaciations. Quat. Sci. Rev. 22:1561–1568.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 11:446–450.

Eltis D, Walvin J. 1981. The abolition of the atlantic slave trade: Origins and effects in Europe, Africa, and the Americas. Madison (WI): Univ of Wisconsin Pr. 314pp.

Eltis D. 2008. A brief overview of the trans-atlantic slave trade. Voyag. Trans-Atlantic Slave Trade Database, http//www. slavevoyages. org/tast/assessment/essays-intro-01. faces (accessed April 27, 2008).

Eltis D, Richardson D. 2010. The Transatlantic Slave Trade Database., http//www. slavevoyages. org (accessed November 20, 2015).

Ernster L, Schatz G. 1981. Mitochondria: a historical review. J. Cell Biol. 91:227s–255s.

Farooqui AA, Horrocks LA, Farooqui T. 2000. Deacylation and reacylation of neural membrane glycerophospholipids. J. Mol. Neurosci. 14:123–135.

Fernandes V, Alshamali F, Alves M, Costa MD, Pereira JB, Silva NM, Cherni L, Harich N, Cerny V, Soares P, Richards MB, Pereira L. 2012. The Arabian cradle: Mitochondrial relicts of the first steps along the Southern route out of Africa. Am. J. Hum. Genet. 90:347–355.

Fischer A, Gilad Y, Man O, Pääbo S. 2005. Evolution of bitter taste receptors in humans and apes. Mol. Biol. Evol. 22:432–436.

Franceschini N, Fox E, Zhang Z, Edwards TL, Nalls M a., Sung YJ, Tayo BO, Sun Y V., Gottesman O, Adeyemo A, Johnson AD, Young JH, Rice K, Duan Q, Chen F, Li Y, Tang H, Fornage M, Keene KL, et al. 2013. Genome-wide association analysis of blood-pressure traits in african-ancestry individuals reveals common associated genes in African and Non-African populations. Am. J. Hum. Genet. 93:545–554.

Freese E. 1959. The specific mutagenic effect of base analogues on phage T4. J. Mol. Biol. 1:87–105.

Fridman C, Gonzalez RS, Pereira AC, Cardena MMSG. 2014. Haplotype diversity in mitochondrial DNA hypervariable region in a population of southeastern Brazil. Int. J. Legal Med. 128:589–593.

Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prufer K, de Filippo C, Meyer M, Zwyns N, Salazar-Garcia DC, Kuzmin Y V, Keates SG, Kosintsev PA, Razhev DI, Richards MP, Peristov N V, et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514:445–449.

Geggus D. 2001. The French slave trade: an overview. William Mary Q. 58:119–138.

Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI, Bowden DW, Langefeld CD, Oleksyk TK, Uscinski Knob AL, Bernhardy AJ, Hicks PJ, Nelson GW, Vanhollebeke B, Winkler CA, Kopp JB, Pays E, Pollak MR. 2010. Association of

Trypanolytic ApoL1 Variants with Kidney Disease in African Americans. Science. 329:841–845.

Gerbault P, Liebert A, Itan Y, Powell A, Currat M, Burger J, Swallow DM, Thomas MG. 2011. Evolution of lactase persistence: an example of human niche construction. Philos. Trans. R. Soc. B Biol. Sci. 366:863–877.

Giles RE, Blanc H, Cann HM, Wallace DC. 1980. Maternal inheritance of human mitochondrial DNA. Proc. Natl. Acad. Sci. 77:6715–6719.

Goebel T, Waters MR, O'Rourke DH. 2008. The late Pleistocene dispersal of modern humans in the Americas. Science 319:1497–1502.

Goedecke JH, Utzschneider K, Faulenbach M V., Rizzo M, Berneis K, Spinas G a., Dave J a., Levitt NS, Lambert E V., Olsson T, Kahn SE. 2010. Ethnic differences in serum lipoproteins and their determinants in South African women. Metabolism. 59:1341–1350.

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson S a, O'connell RJ, Agan BK, Ahuja SKSS, Bologna R, et al. 2005. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307:1434–1440.

Gordon M. 1989. Slavery in the Arab world. New York: New Amsterdam Books. 272 pp.

Gray MW. 1993. Origin and evolution of organelle genomes. Curr. Opin. Genet. Dev. 3:884–890.

Grugni V, Battaglia V, Hooshiar Kashani B, Parolo S, Al-Zahery N, Achilli A, Olivieri A, Gandini F, Houshmand M, Sanati MH, Torroni A, Semino O. 2012. Ancient migratory events in the Middle East: new clues from the Y-chromosome variation of modern Iranians. PLoS One 7:e41252.

Gubler DJ. 2004. Cities spawn epidemic dengue viruses. Nat. Med. 10:129–130.

Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, Ritchie GRS, Xue Y, Asimit J, Nsubuga RN, Young EH, Pomilla C, Kivinen K, Rockett K, Kamali A, et al. 2015. The African Genome Variation Project shapes medical genetics in Africa. Nature 517:327–332.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, Fu Q, Mittnik A, Bánffy E, Economou C, Francken M, Friederich S, Pena RG, Hallgren F, Khartanovich V, Khokhlov A, Kunst M, Kuznetsov P, Meller H, Mochalov O, Moiseyev V, Nicklisch N, Pichler SL, Risch R, Rojo Guerra MA, Roth C, Szécsényi-Nagy A, Wahl J, Meyer M, Krause J, Brown D, Anthony D, Cooper A, Alt KW, Reich D. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 522:207-211.

Harich N, Costa MD, Fernandes V, Kandil M, Pereira JB, Silva NM, Pereira L. 2010. The trans-Saharan slave trade - clues from interpolation analyses and high-resolution characterization of mitochondrial DNA lineages. BMC Evol. Biol. 10:138.

Harris JB, LaRocque RC, Chowdhury F, Khan AI, Logvinenko T, Faruque ASG, Ryan ET, Qadri F, Calderwood SB. 2008. Susceptibility to Vibrio cholerae infection in a cohort of household contacts of patients with cholera in Bangladesh. PLoS Negl. Trop. Dis. 2:e221.

Hassan HY, Underhill PA, Cavalli-Sforza LL, Ibrahim ME. 2008. Y-chromosome variation among Sudanese: restricted gene flow, concordance with language, geography, and history. Am. J. Phys. Anthropol. 137:316–323.

Hastie T, Tibshirani R, Narasimhan B, Chu G. 2012. Impute: Imputation for microarray data. R Package version 1.

Heinz T, Alvarez-Iglesias V, Pardo-Seco J, Taboada-Echalar P, Gómez-Carballa A, Torres-Balanza A, Rocabado O, Carracedo A, Vullo C, Salas A. 2013. Ancestry analysis reveals a predominant Native American component with moderate European admixture in Bolivians. Forensic Sci. Int. Genet. 7:537–542.

Heinz-Erian P, Müller T, Krabichler B, Schranz M, Becker C, Rüschendorf F, Nürnberg P, Rossier B, Booth IW, Holmberg C, Wijmenga C, Grigelioniene G, Kneepkens CMF, Rosipal S, Mistrik M, Kappler M, Michaud L, Dóczy LC, Siu VM, et al. 2008. Mutations in SPINT2 cause a syndromic form of congenital sodium diarrhea. Am. J. Hum. Genet. 84:188–196.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. Science 343:747–751.

Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, Rodríguez-Botigué L, Ramachandran S, Hon L, Brisbin A, Lin AA, Underhill PA, Comas D, Kidd KK, Norman PJ, Parham P, Bustamante CD, Mountain JL, Feldman MW. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc. Natl. Acad. Sci. U. S. A. 108:5154–5162.

Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, Fadhlaoui-Zid K, Zalloua PA, Moreno-Estrada A, Bertranpetit J, Bustamante CD, Comas D. 2012. Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet. 8:e1002397.

Hewitt G. 2000. The genetic legacy of the Quaternary ice ages. Nature 405:907–913.

Hodgson JA, Mulligan CJ, Al-Meeri A, Raaum RL. 2014. Early back-to-Africa migration into the Horn of Africa. PLoS Genet. 10:e1004393.

Howes RE, Patil AP, Piel FB, Nyangiri O a, Kabaria CW, Gething PW, Zimmerman P a, Barnadas C, Beall CM, Gebremedhin A, Ménard D, Williams TN, Weatherall DJ, Hay SI. 2011. The global distribution of the Duffy blood group. Nat. Commun. 2:266.

Jablonski NG, Chaplin G. 2000. The evolution of human skin coloration. J. Hum. Evol. 39:57–106.

Janeway CA, Travers P, Walport MJ, Shlomchik MJ. 2001. Immunobiology: the immune system in health and disease. London: Churchill Livingstone. 848pp.

Jha V, Garcia-Garcia G, Iseki K, Li Z, Naicker S, Plattner B, Saran R, Wang AY-M, Yang C-W. 2013. Chronic kidney disease: global dimension and perspectives. Lancet 382:260–272.

Jobling M, Hurles M, Tyler-Smith C. 2013. Human evolutionary genetics: origins, peoples & disease. Garland Science. 650pp.

Joseph SB, Castrillo A, Laffitte BA, Mangelsdorf DJ, Tontonoz P. 2003. Reciprocal regulation of inflammation and lipid metabolism by liver X receptors. Nat. Med. 9:213–219.

Karlsson EK, Kwiatkowski DP, Sabeti PC. 2014. Natural selection and infectious disease in human populations. Nat. Rev. Genet. 15:379–393.

Khor CC, Chau TNB, Pang J, Davila S, Long HT, Ong RTH, Dunstan SJ, Wills B, Farrar J, Van Tram T, Gan TT, Binh NTN, Tri LT, Lien LB, Tuan NM, Tham NTH, Lanh MN, Nguyet NM, Hieu NT, et al. 2011. Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1. Nat. Genet. 43:1139–1141.

Klein HS. 1999. The Atlantic Slave Trade. Cambridge: Cambridge University Press. 256pp.

Ko W-Y, Rajan P, Gomez F, Scheinfeldt L, An P, Winkler CA, Froment A, Nyambo TB, Omar SA, Wambebe C, Ranciaro A, Hirbo JB, Tishkoff SA. 2013. Identifying Darwinian selection acting on different human APOL1 variants among diverse African populations. Am. J. Hum. Genet. 93:54–66.

Kuhn PC, Horimoto ARVR, Sanches JM, Vieira Filho JPB, Franco L, Fabbro AD, Franco LJ, Pereira AC, Moises RS. 2012. Genome-wide analysis in Brazilian Xavante Indians reveals low degree of admixture. PLoS One 7:e42702.

Kulkarni H, Marconi VC, He W, Landrum ML, Okulicz JF, Delmar J, Kazandjian D, Castiblanco J, Ahuja SS, Wright EJ, Weiss RA., Clark RA., Dolan MJ, Ahuja SK. 2009. The Duffy-null state is associated with a survival advantage in leukopenic HIV-infected persons of African ancestry. Blood 114:2783–2792.

Kwiatkowski DP. 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. Am. J. Hum. Genet. 77:171–192.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, Berger B, Economou C, Bollongino R, Fu Q, Bos KI, Nordenfelt S, Li H, de Filippo C, Prufer K, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409–413.

Ledda M, Kutalik Z, Souza Destito MC, Souza MM, Cirillo CA, Zamboni A, Martin N, Morya E, Sameshima K, Beckmann JS, le Coutre J, Bergmann S, Genick UK, Destito MCS, Souza MM, Cirillo CA, Zamboni A, Martin N, Morya E, et al. 2014. GWAS of human bitter taste perception identifies new loci and reveals additional complexity of bitter taste genetics. Hum. Mol. Genet. 23:259–267.

Lee RJ, Cohen NA. 2014. Taste receptors in innate immunity. Cell. Mol. Life Sci. 72:217–236.

Li D, Zhang J. 2014. Diet shapes the evolution of the vertebrate bitter taste receptor gene repertoire. Mol. Biol. Evol. 31:303–309.

Lin Y, Sun Z. 2010. Current views on type 2 diabetes. J. Endocrinol. 204:1–11.

Lind JM, Hutcheson-Dilks HB, Williams SM, Moore JH, Essex M, Ruiz-Pesini E, Wallace DC, Tishkoff SA, O'Brien SJ, Smith MW. 2007. Elevated male European and female African contributions to the genomes of African American individuals. Hum. Genet. 120:713–722.

Llorente MG, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, Curtis MC, Stock JT, Coltorti M, Pieruccini P, others. 2015. Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. Science. 350:820-822.

Loke H, Bethell D, Phuong CXT, Day N, White N, Farrar J, Hill A. 2002. Susceptibility to dengue hemorrhagic fever in Vietnam: Evidence of an association with variation in the vitamin D receptor and FCγ receptor IIA genes. Am. J. Trop. Med. Hyg. 67:102–106.

Long J, Edwards T, Signorello LB, Cai Q, Zheng W, Shu X-O, Blot WJ. 2012. Evaluation of genome-wide association study-identified type 2 diabetes loci in African Americans. Am. J. Epidemiol. 176:995–1001.

Ma F, Liu S-Y, Razani B, Arora N, Li B, Kagechika H, Tontonoz P, Núñez V, Ricote M, Cheng G. 2014. Retinoid X receptor α attenuates host antiviral response by suppressing type I interferon. Nat. Commun. 5:5494.

Manning P. 1990. Slavery and African life: occidental, oriental, and African slave trades. Cambridge: Cambridge University Press. 252 pp.

Martel-Jantin C, Filippone C, Tortevoye P, Afonso P V, Betsem E, Descorps-Declere S, Nicol JTJ, Touzé A, Coursaget P, Crouzat M, Berthet N, Cassar O, Gessain A. 2014. Molecular

epidemiology of merkel cell polyomavirus: evidence for geographically related variant genotypes. J. Clin. Microbiol. 52:1687–1690.

Masel J. 2011. Genetic drift. Curr. Biol. 21:R837–R838.

Maskarinec G, Grandinetti A, Matsuura G, Sharma S, Mau M, Henderson BE, Kolonel LN. 2009. Diabetes prevalence and body mass index differ by ethnicity: the Multiethnic Cohort. Ethn. Dis. 19:49–55.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Llamas B, Pickrell J, Meller H, Rojo Guerra MA, Krause J, Anthony D, Brown D, Lalueza Fox C, Cooper A, Alt KW, Haak W, Patterson N, Reich D. 2015. Eight thousand years of natural selection in Europe. bioRxiv, 016477.

McCulloch SD, Kunkel TA. 2008. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. Cell Res. 18:148–161.

McGilvray ID, Serghides L, Kapus A, Rotstein OD, Kain KC. 2000. Nonopsonic monocyte/macrophage phagocytosis of Plasmodium falciparum-parasitized erythrocytes: a role for CD36 in malarial clearance. Blood 96:3231–3240.

Melmed S, Conn PM. 2005. Endocrinology: Basic and clinical principles. Totowa (NJ): Humana Press. 440 pp.

Mendizabal I, Sandoval K, Berniell-Lee G, Calafell F, Salas A, Martínez-Fuentes A, Comas D. 2008. Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba. BMC Evol. Biol. 8:213.

Mendizabal I, Marigorta UM, Lao O, Comas D. 2012. Adaptive evolution of loci covarying with the human African Pygmy phenotype. Hum. Genet. 131:1305–1317.

Messer PW, Petrov DA. 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol. Evol. 28:659–669.

Michaux JR, Libois R, Paradis E, Filippucci M-G. 2004. Phylogeographic history of the yellow-necked fieldmouse (Apodemus flavicollis) in Europe and in the Near and Middle East. Mol. Phylogenet. Evol. 32:788–798.

Migliano AB, Romero IG, Metspalu M, Leavesley M, Pagani L, Antao T, Huang D-W, Sherman BT, Siddle K, Scholes C, Hudjashov G, Kaitokai E, Babalu A, Belatti M, Cagan A, Hopkinshaw B, Shaw C, Nelis M, Metspalu E, et al. 2013. Evolution of the pygmy phenotype: evidence of positive selection from genome-wide scans in African, Asian, and Melanesian pygmies. Hum. Biol. 85:251–284.

Mohammed-Ali AS, Khabir AM. 2003. The Wavy Line and the Dotted Wavy Line Pottery in the Prehistory of the Central Nile and the Sahara-Sahel Belt. African Archaeol. Rev. 20:25–58.

Montana G, Hoggart C. 2007. Statistical software for gene mapping by admixture linkage disequilibrium. Brief. Bioinform. 8:393–395.

Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D. 2011. The History of African Gene Flow into Southern Europeans, Levantines, and Jews. PLoS Genet 7:e1001373.

Mörner M. 1967. Race mixture in the history of Latin America. Boston (MA): Little Brown & Company. 178pp.

Motojima K, Passilly P, Peters JM, Gonzalez FJ, Latruffe N. 1998. Expression of putative fatty acid transporter genes are regulated by peroxisome proliferator-activated receptor alpha and gamma activators in a tissue- and inducer-specific manner. J. Biol. Chem. 273:16710–16714.

Myers S, Spencer CCA, Auton A, Bottolo L, Freeman C, Donnelly P, McVean G. 2006. The distribution and causes of meiotic recombination in the human genome. Biochem. Soc. Trans. 34:526–530.

Myers S, Freeman C, Auton A, Donnelly P, McVean G. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. Nat. Genet. 40:1124–1129.

Naar AM, Najafi-Shoushtari SH. 2013. Methods targeting mir-128 for regulating cholesterol/lipid metabolism. U.S. Patent Application No. 13/979,428.

Neumann K. 2003. The late emergence of agriculture in sub-Saharan Africa: archaeobotanical evidence and ecological considerations. In Neumann K, Butler A, Kahlheber S (Ed) Food, fuel and fields: Progress in African archaeobotany.Köln: Heinrich-Barth-Institute. Pp71–92.

Newman JL. 1995. The peopling of Africa: a geographic interpretation. New Haven: Yale University Press. 235pp.

Ng MC, Shriner D, Chen BH, Li J, Chen WM, Guo X, Liu J, Bielinski SJ, Yanek LR, Nalls MA, Comeau ME, Rasmussen-Torvik LJ, Jensen RA, Evans DS, Sun Y V, An P, Patel SR, Lu Y, Long J, et al. 2014. Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. PLoS Genet 10:e1004517.

Nickel RG, Willadsen SA, Freidhoff LR, Huang S-K, Caraballo L, Naidu RP, Levett P, Blumenthal M, Banks-Schlegel S, Bleecker E, Beaty T, Ober C, Barnes KC. 1999. Determination of Duffy genotypes in three populations of African descent using PCR and sequence-specific oligonucleotides. Hum. Immunol. 60:738–742.

Normile D. 2013. Surprising New Dengue Virus Throws A Spanner in Disease Control Efforts. Science. 342:2013.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR. 2008. Genes mirror geography within Europe. Nature 456:98–101.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 27:29–34.

Ojodu J, Hulihan MM, Pope SN, Grant AM. 2014. Incidence of sickle cell trait-United States, 2010. MMWR. Morb. Mortal. Wkly. Rep. 63:1155–1158.

Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, Al-Zahery N, Scozzari R, Cruciani F, Behar DM, Dugoujon J-M, Coudray C, Santachiara-Benerecetti a S, Semino O, Bandelt H-J, Torroni A. 2006. The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. Science 314:1767–70.

Ong KL, Cheung BMY, Man YB, Lau CP, Lam KSL. 2007. Prevalence, awareness, treatment, and control of hypertension among United States adults 1999-2004. Hypertension 49:69–75.

Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Gallego Romero I, Ayub Q, Mehdi SQ, Thomas MG, Luiselli D, Bekele E, Bradman N, Balding DJ, Tyler-Smith C. 2012. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am. J. Hum. Genet. 91:83–96.

Pala M, Olivieri A, Achilli A, Accetturo M, Metspalu E, Reidla M, Tamm E, Karmin M, Reisberg T, Kashani BH, Perego U a., Carossa V, Gandini F, Pereira JB, Soares P, Angerhofer N,

Rychkov S, Al-Zahery N, Carelli V, et al. 2012. Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia. Am. J. Hum. Genet. 90:915–924.

Palmer ND, McDonough CW, Hicks PJ, Roh BH, Wing MR, An SS, Hester JM, Cooke JN, Bostrom M a, Rudock ME, Talbert ME, Lewis JP, Ferrara A, Lu L, Ziegler JT, Sale MM, Divers J, Shriner D, Adeyemo A, et al. 2012. A genome-wide association search for type 2 diabetes genes in African Americans. PLoS One 7:e29202.

Parsa A, Kao WHL, Xie D, Astor BC, Li M, Hsu C, Feldman HI, Parekh RS, Kusek JW, Greene TH, Fink JC, Anderson AH, Choi MJ, Wright JT, Lash JP, Freedman BI, Ojo A, Winkler C a, Raj DS, et al. 2013. APOL1 risk variants, race, and progression of chronic kidney disease. N. Engl. J. Med. 369:2183–2196.

Patin E, Laval G, Barreiro LB, Salas A, Semino O, Santachiara-Benerecetti S, Kidd KK, Kidd JR, Van der Veen L, Hombert J-M, Gessain A, Froment A, Bahuchet S, Heyer E, Quintana-Murci L. 2009. Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. PLoS Genet. 5:e1000448.

Patin E, Siddle KJ, Laval G, Quach H, Harmant C, Becker N, Froment A, Régnault B, Lemée L, Gravel S, others. 2014. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nat. Commun. 5:3163.

Pena SDJ, Di Pietro G, Fuchshuber-Moraes M, Genro JP, Hutz MH, Kehdy F de SG, Kohlrausch F, Magno LAV, Montenegro RC, Moraes MO, de Moraes MEA, de Moraes MR, Ojopi EB, Perini JA, Racciopi C, Ribeiro-Dos-Santos AKC, Rios-Santos F, Romano-Silva MA, Sortica VA, et al. 2011. The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. PLoS One 6:e17063.

Pereira L, Richards M, Goios A, Alonso A, Albarrán C, Garcia O, Behar DM, Gölge M, Hatina J, Al-Gazali L, Bradley DG, Macaulay V, Amorim A. 2005. High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. Genome Res. 15:19–24.

Pérez-Morga D, Vanhollebeke B, Paturiaux-Hanocq F, Nolan DP, Lins L, Homblé F, Vanhamme L, Tebabi P, Pays A, Poelvoorde P, Jacquet A, Brasseur R, Pays E. 2005. Apolipoprotein L-I promotes trypanosome lysis by forming pores in lysosomal membranes. Science 309:469–472.

Petit RJ, Raynaud D, Basile I, Chappellaz J, Ritz C, Delmotte M, Legrand M, Lorius C, Pe L. 1999. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. Nature 399:429–413.

Phillipson DW. 2010. The first millennium BC in the highlands of Northern Ethiopia and South-central Eritrea: A reassessment of cultural and political development. African Archaeol. Rev. 26:257–274.

Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D. 2014. Ancient west Eurasian ancestry in southern and eastern Africa. Proc. Natl. Acad. Sci. U. S. A. 111:2632–2637.

Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, Kure B, Mpoloka SW, Nakagawa H, Naumann C, Lipson M, Loh P-R, Lachance J, Mountain J, Bustamante CD, Berger B, Tishkoff SA, Henn BM, Stoneking M, et al. 2012. The genetic prehistory of southern Africa. Nat. Commun. 3:1143.

Plaza S, Salas A, Calafell F, Corte-Real F, Bertranpetit J, Carracedo Á, Comas D. 2004. Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. Hum. Genet. 115:439–447.

Podgorná E, Soares P, Pereira L, Cerný V. 2013. The genetic impact of the Lake Chad basin population in North Africa as documented by mitochondrial diversity and internal variation of the L3e5 haplogroup. Ann. Hum. Genet. 77:513–523.

Prentki M, Madiraju SRM. 2008. Glycerolipid metabolism and signaling in health and disease. Endocr. Rev. 29:647–676.

Price TD, Tiesler V, Burton JH. 2006. Early African Diaspora in colonial Campeche, Mexico: Strontium isotopic evidence. Am. J. Phys. Anthropol. 130:485–490.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Pritchard JK, Pickrell JK, Coop G. 2010. The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Curr. Biol. 20:R208–R215.

Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F. 2005. Pathogen-driven selection and worldwide HLA class I diversity. Curr. Biol. 15:1022–1027.

Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. 2010. LocusZoom: Regional visualization of genome-wide association scan results. Bioinformatics 26:2336–2337.

Ramsuran V, Kulkarni H, He W, Mlisana K, Wright EJ, Werner L, Castiblanco J, Dhanda R, Le T, Dolan MJ, Guan W, Weiss R a., Clark R a., Abdool Karim SS, Ahuja SK, Ndung'U T. 2011. Duffy-null-associated low neutrophil counts influence HIV-1 susceptibility in high-risk South African black women. Clin. Infect. Dis. 52:1248–1256.

Ranciaro A, Campbell MC, Hirbo JB, Ko W-Y, Froment A, Anagnostou P, Kotze MJ, Ibrahim M, Nyambo T, Omar SA, Tishkoff SA. 2014. Genetic origins of lactase persistence and the spread of pastoralism in Africa. Am. J. Hum. Genet. 94:496–510.

Ranjit S, Kissoon N. 2011. Dengue hemorrhagic fever and shock syndromes. Pediatr. Crit. Care Med. 12:90–100.

Rastogi S. 2011. The black population: 2010. US Department of Commerce, Economics and Statistics Administration, US Census Bureau.

Rawley JA, Behrendt SD. 2005. The transatlantic slave trade: a history. Lincoln (NE): University of Nebraska Press. 464 pp.

Rees DC, Williams TN, Gladwin MT. 2010. Sickle-cell disease. Lancet 376:2018–2031.

Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altshuler D. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. Nat. Genet. 32:135–142.

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra M V., Rojas W, Duque C, Mesa N, García LF, Triana O, Ruiz-Linares A, et al. 2012. Reconstructing Native American population history. Nature 488:370–374.

Reiter P. 2010. Yellow fever and dengue: a threat to Europe? Euro Surveill. 15:19509.

Richards M, Rengo C, Cruciani F, Gratrix F, Wilson JF, Scozzari R, Macaulay V, Torroni A. 2003. Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations. Am. J. Hum. Genet. 72:1058–1064.

Richards M, Bandelt H-J, Kivisild T, Oppenheimer S. 2006. A model for the dispersal of modern humans out of Africa. In Bandelt, H. J., Richards, M., & Macaulay, V. : Human Mitochondrial DNA and the Evolution of Homo sapiens. Springer. Pp. 225–265.

Rodenhuis-Zybert IA, Wilschut J, Smit JM. 2010. Dengue virus life cycle: viral and host factors modulating infectivity. Cell. Mol. Life Sci. 67:2773–2786.

Rodríguez F, Hammer S, Pérez T, Suchentrunk F, Lorenzini R, Michallet J, Martinkova N, Albornoz J, Domínguez A. 2009. Cytochrome b phylogeography of chamois (Rupicapra spp.). Population contractions, expansions and hybridizations governed the diversification of the genus. J. Hered. 100:47–55.

Rose JI, Petraglia MD. 2010. Tracking the origin and evolution of human populations in Arabia. Amsterdam: Springer. 312 pp.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll S a, Gaudet R, Schaffner SF, Lander ES, Frazer K a, Ballinger DG, Cox DR, Hinds D a, Stuve LL, Gibbs R a, Belmont JW, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918.

Salas A, Jaime JC, Alvarez-Iglesias V, Carracedo A. 2008. Gender bias in the multiethnic genetic composition of central Argentina. J. Hum. Genet. 53:662–674.

Saraiva RM, Hare JM. 2006. Nitric oxide signaling in the cardiovascular system: implications for heart failure. Curr. Opin. Cardiol. 21:221–228.

Schroeder H, O'Connell TC, Evans JA, Shuler KA, Hedges REM. 2009. Trans-atlantic slavery: Isotopic evidence for forced migration to Barbados. Am. J. Phys. Anthropol. 139:547–557.

Schroeder H, Ávila-Arcos MC, Malaspinas A-S, Poznik GD, Sandoval-Velasco M, Carpenter ML, Moreno-Mayar JV, Sikora M, Johnson PLF, Allentoft ME, Samaniego JA, Haviser JB, Dee MW, Stafford TW, Salas A, Orlando L, Willerslev E, Bustamante CD, Gilbert MTP. 2015. Genome-wide ancestry of 17th-century enslaved Africans from the Caribbean. Proc. Natl. Acad. Sci. U. S. A. 112:3669-3673.

Schulz A, Israel B, Williams D, Parker E, Becker A, James S. 2000. Social inequalities, stressors and self reported health status among African American and white women in the Detroit metropolitan area. Soc. Sci. Med. 51:1639–1653.

Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas A-S, Manica A, Moltke I, Albrechtsen A, Ko A, Margaryan A, Moiseyev V. 2014. Genomic structure in Europeans dating back at least 36,200 years. Science. 346:1113–1118.

Semino O, Santachiara-Benerecetti AS, Falaschi F, Cavalli-Sforza LL, Underhill PA. 2002. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. Am. J. Hum. Genet. 70:265–268.

Sheiner E, Levy A, Yerushalmi R, Katz M. 2004. Beta-thalassemia minor during pregnancy. Obstet. Gynecol. 103:1273–1277.

Silva DA, Carvalho E, Costa G, Tavares L, Amorim A, Gusmão L. 2006. Y-chromosome genetic variation in Rio de Janeiro population. Am. J. Hum. Biol. 18:829–837.

Silva LK, Blanton RE, Parrado AR, Melo PS, Morato VG, Reis EA, Dias JP, Castro JM, Vasconcelos PF, Goddard KA, Barreto ML, Reis MG, Teixeira MG. 2010. Dengue hemorrhagic fever is associated with polymorphisms in JAK1. Eur J Hum Genet 18:1221–1227.

Sinha RP, Häder D-P. 2002. UV-induced DNA damage and repair: a review. Photochem. Photobiol. Sci. 1:225–236.

Smith LS. 1999. MacMillan Encyclopedia of World Slavery. Ref. User Serv. Q. 38:418.

Soares P, Trejaut JA, Loo JH, Hill C, Mormina M, Lee CL, Chen YM, Hudjashov G, Forster P, MacAulay V, Bulbeck D, Oppenheimer S, Lin M, Richards MB. 2008. Climate change and postglacial human dispersals in Southeast Asia. Mol. Biol. Evol. 25:1209–1218.

Sommer R, Benecke N. 2005. Late-Pleistocene and early Holocene history of the canid fauna of Europe (Canidae). Mamm. Biol. 70:227–241.

Sommer RS, Nadachowski A. 2006. Glacial refugia of mammals in Europe: Evidence from fossil records. Mamm. Rev. 36:251–265.

Steinberg MH, Forget BG, Higgs DR, Weatherall DJ. 2009. Disorders of hemoglobin: genetics, pathophysiology, and clinical management. Cambridge: Cambridge University Press. 846 pp.

Stokes MJ, Murakami Y, Maeda Y, Kinoshita T, Morita YS. 2014. New insights into the functions of PIGF, a protein involved in the ethanolamine phosphate transfer steps of glycosylphosphatidylinositol biosynthesis. Biochem. J. 463:249–256.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Eichler EE. 2010. Diversity of human copy number variation and multicopy genes. Science 330:641–646.

Tell GS, Hylander B, Craven TE, Burkart J. 1996. Racial Differences in the Incidence of End-Stage Renal Disease. Ethn. Health 1:21–31.

Teslovich, T. Musunuru, K. Smith a. E Al. 2010. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466:707–713.

Thomas H. 1997. The Slave Trade: The Story of the Atlantic Slave Trade: 1440-1870. New York: Simon & Schuster. 912 pp.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. 39:31–40.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O, et al. 2009. The genetic structure and history of Africans and African Americans. Science 324:1035–1044.

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ. 2006. Harvesting the fruit of the human mtDNA tree. Trends Genet. 22:339–345.

Vanhamme L, Paturiaux-Hanocq F, Poelvoorde P, Nolan DP, Lins L, Van Den Abbeele J, Pays A, Tebabi P, Van Xong H, Jacquet A, Moguilevsky N, Dieu M, Kane JP, De Baetselier P, Brasseur R, Pays E. 2003. Apolipoprotein L-I is the trypanosome lytic factor of human serum. Nature 422:83–87.

Vesa TH, Marteau P, Korpela R. 2000. Lactose intolerance. J. Am. Coll. Nutr. 19:165S–175S.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLOS Biol. 4:e72.

Wagh K, Bhatia A, Alexe G, Reddy A, Ravikumar V, Seiler M, Boemo M, Yao M, Cronk L, Naqvi A, Ganesan S, Levine AJ, Bhanot G. 2012. Lactase persistence and lipid pathway selection in the Maasai. PLoS One 7:e44751.

Wang X, Thomas SD, Zhang J. 2004. Relaxation of selective constraint and loss of function in the evolution of human bitter taste receptor genes. Hum. Mol. Genet. 13:2671–2678.

Wang S, Ray N, Rojas W, Parra M V, Bedoya G, Gallo C, Poletti G, Mazzotti G, Hill K, Hurtado AM, Camrena B, Nicolini H, Klitz W, Barrantes R, Molina JA, Freimer NB, Bortolini MC, Salzano FM, Petzl-Erler ML, Tsuneto LT, Dipierri JE, Alfaro EL, Bailliet G, Bianchi NO, Llop E, Rothhammer F, Excoffier L, Ruiz-Linares A. 2008. Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genet. 4:e1000037.

Wei W, Ayub Q, Chen Y, McCarthy S, Hou Y, Carbone I, Xue Y, Tyler-Smith C. 2013. A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res. 23:388–395.

Wendorf F, Schild R, Close AE. 1984. Cattle-keepers of the eastern Sahara: the neolithic of Bir Kiseiba. Department of Anthropology, Institute for the Study of Earth and Man, Dallas(TX): Southern Methodist University. 452 pp.

WHO 2009. Dengue: guidelines for diagnosis, treatment, prevention, and control. Spec. Program. Res. Train. Trop. Dis.:147.

WHO 2012. Global Strategy for Dengue Prevention and Control 2012–2020. World Heal. Organisation, Geneva, Switz.:35.

Winkler C a, Nelson GW, Smith MW. 2010. Admixture mapping comes of age. Annu. Rev. Genomics Hum. Genet. 11:65–89.

Wintrobe MM, Greer JP. 2009. Wintrobe's clinical hematology. Philadelphia (PA): Lippincott Williams & Wilkins. 2312 pp.

Wood ET, Stover D a, Ehret C, Destro-Bisol G, Spedini G, McLeod H, Louie L, Bamshad M, Strassmann BI, Soodyall H, Hammer MF. 2005. Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. Eur. J. Hum. Genet. 13:867–876.

Wright S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

Wright S. 1943. Isolation by Distance. Genetics 28:114–138.

Wright S. 1950. Genetical structure of populations. Nature 166:247–249.

Wright S. 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution 19:395–420.

Yen C-LE, Stone SJ, Koliwad S, Harris C, Farese R V. 2008. Thematic review series: glycerolipids. DGAT enzymes and triacylglycerol biosynthesis. J. Lipid Res. 49:2283–2301.

Zhu X, Young JH, Fox E, Keating BJ, Franceschini N, Kang S, Tayo B, Adeyemo A, Sun Y V., Li Y, Morrison A, Newton-Cheh C, Liu K, Ganesh SK, Kutlar A, Vasan RS, Dreisbach A, Wyatt S, Polak J, Palmas W, Musani S, Taylor H, Fabsitz R, Townsend RR, Dries D, Glessner J, Chiang CWK, Mosley T, Kardia S, Curb D, Hirschhorn JN, Rotimi C, Reiner A, Eaton C, Rotter JI, Cooper RS, Redline S, Chakravarti A, Levy D. 2011. Combined admixture mapping and association analysis identifies a novel blood pressure genetic locus on 5p13: contributions from the CARe consortium. Hum. Mol. Genet. 20:2285–2295.

Zimmer C. 2001. Evolution: the triumph of an idea. London: Harper Perennial. 528pp.

7 Appendices

7.1 Appendix A – Supplementary Material Paper I.

Extensive admixture and selective pressure across the Sahel Belt.

Genome biology and evolution, *7*(12), 3484-3495.

Due to the large extend of Supplementary Material, only the most relevant supplementary tables and figures are presented in printed version. Complete Supplementary Material is provided in digital format (Supplementary Tables – Paper I).

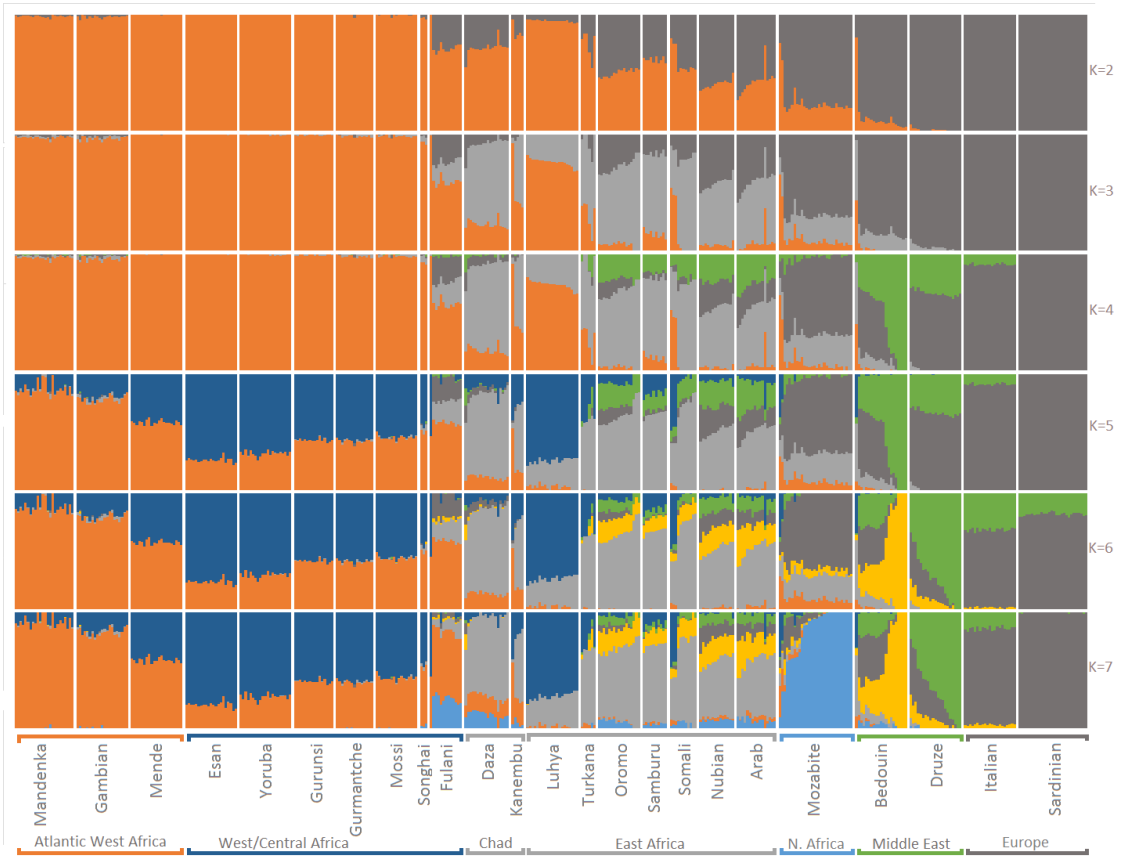

Figure S3 - ADMIXTURE results. Ks between 2 and 7.

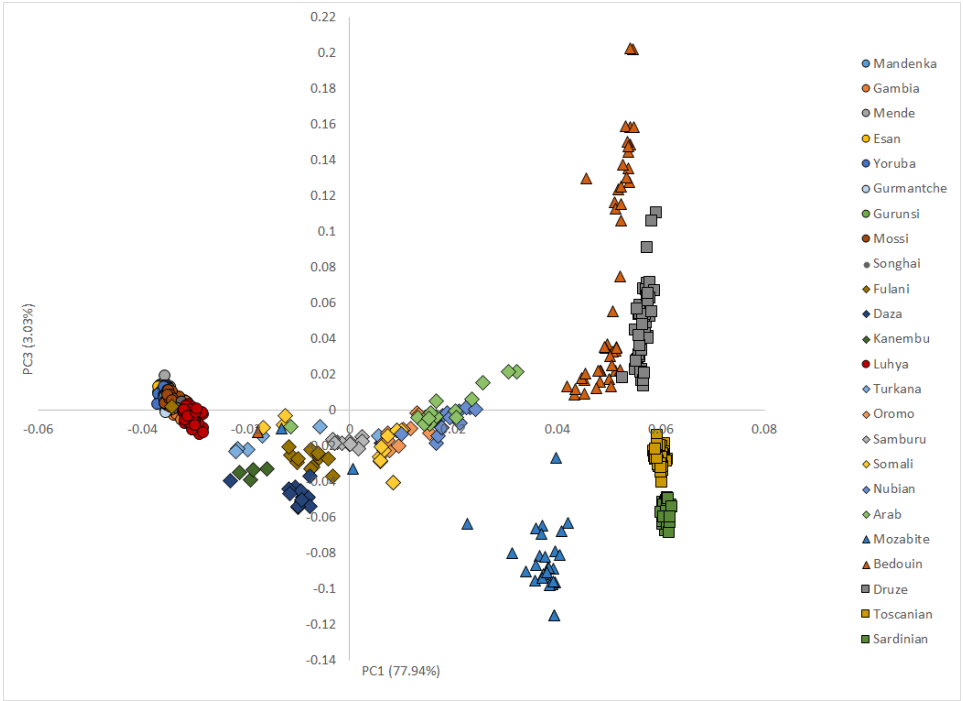

Figure S5 - PC1 versus PC3.

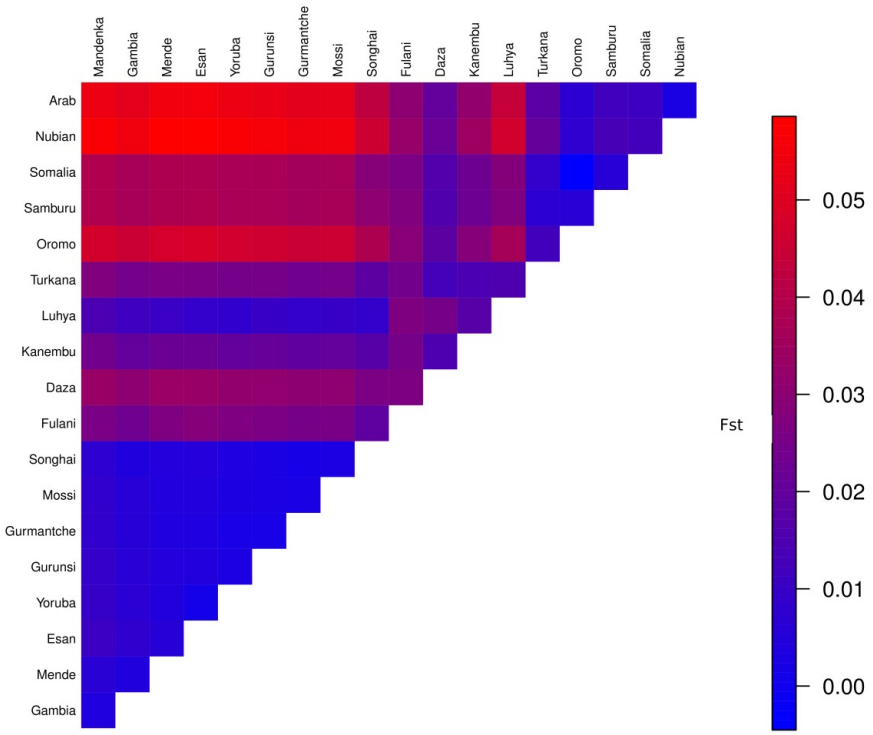

Figure S6 - Heat map for F_{ST} distances between Sahelian populations.

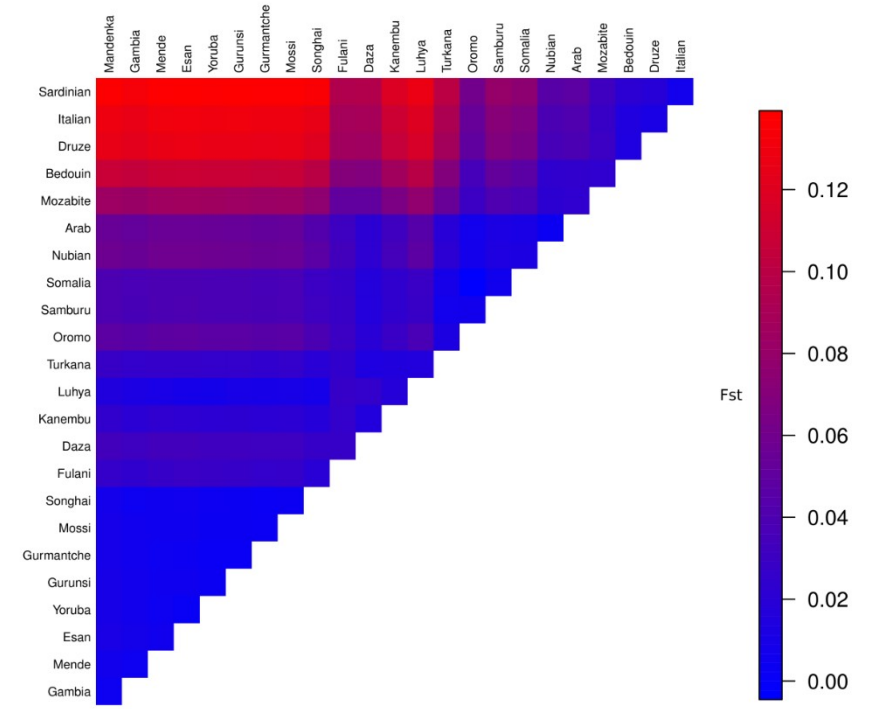

Figure S7 - Heat map for F_{ST} distances between Sahelian and Eurasian populations.

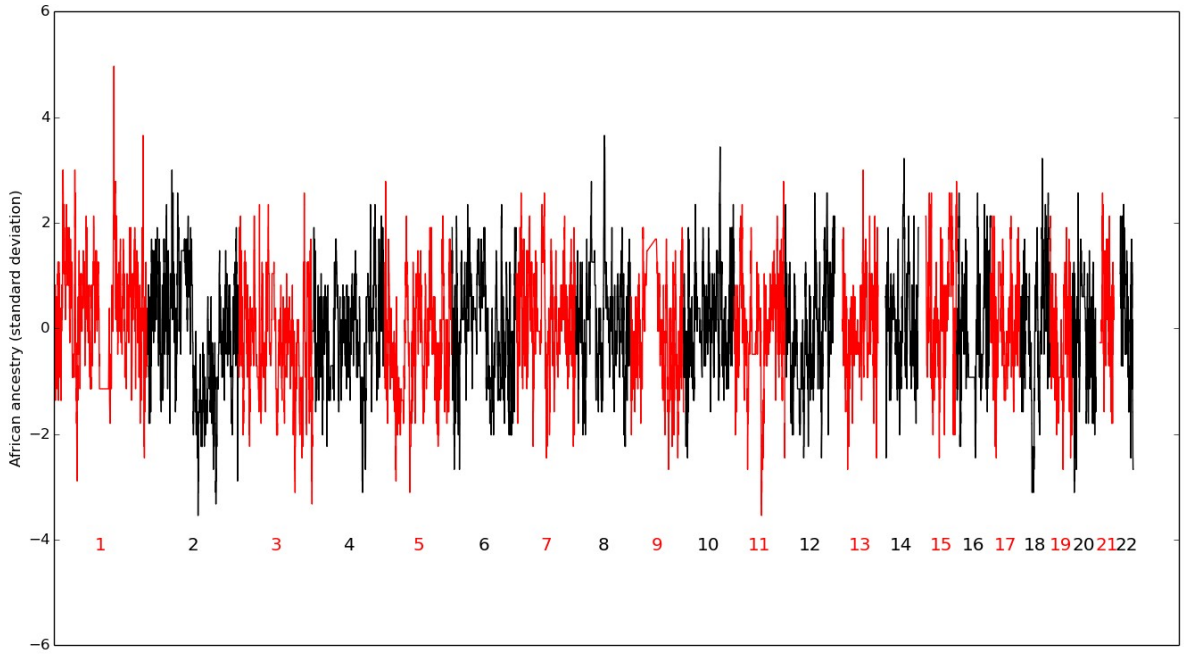

Figure S8 - RFMix results in Arabs and Nubians using Luhya and Italy as parental populations.

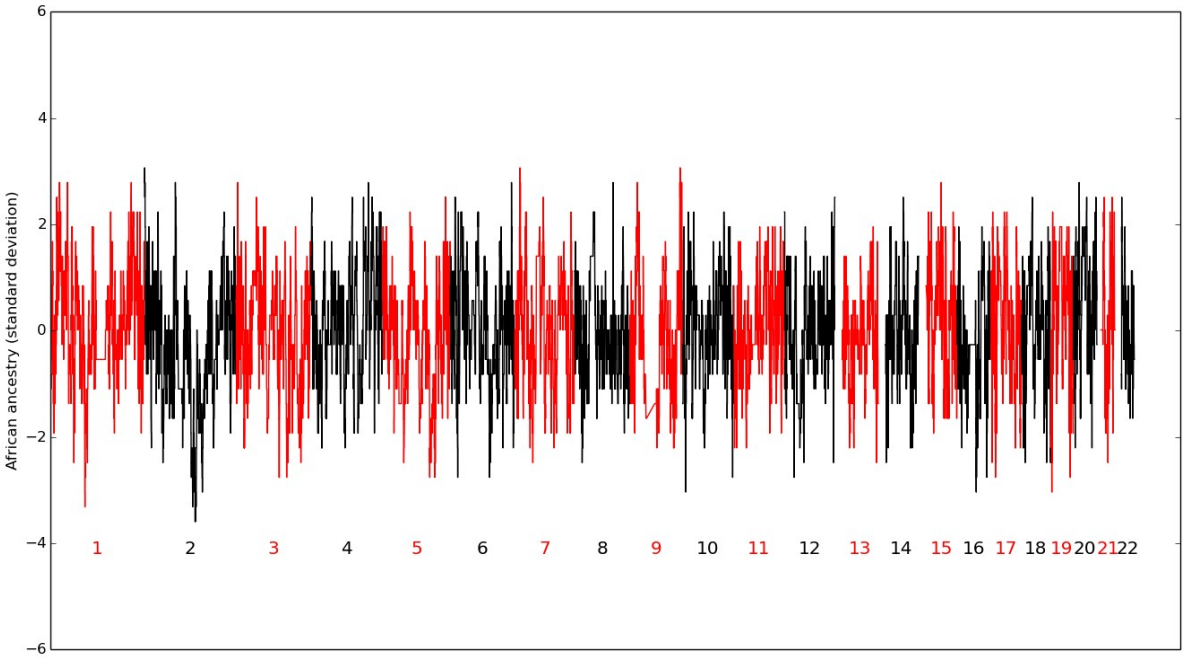

Figure S11 - RFMix results in Oromo using Luhya and Italy as parental populations.

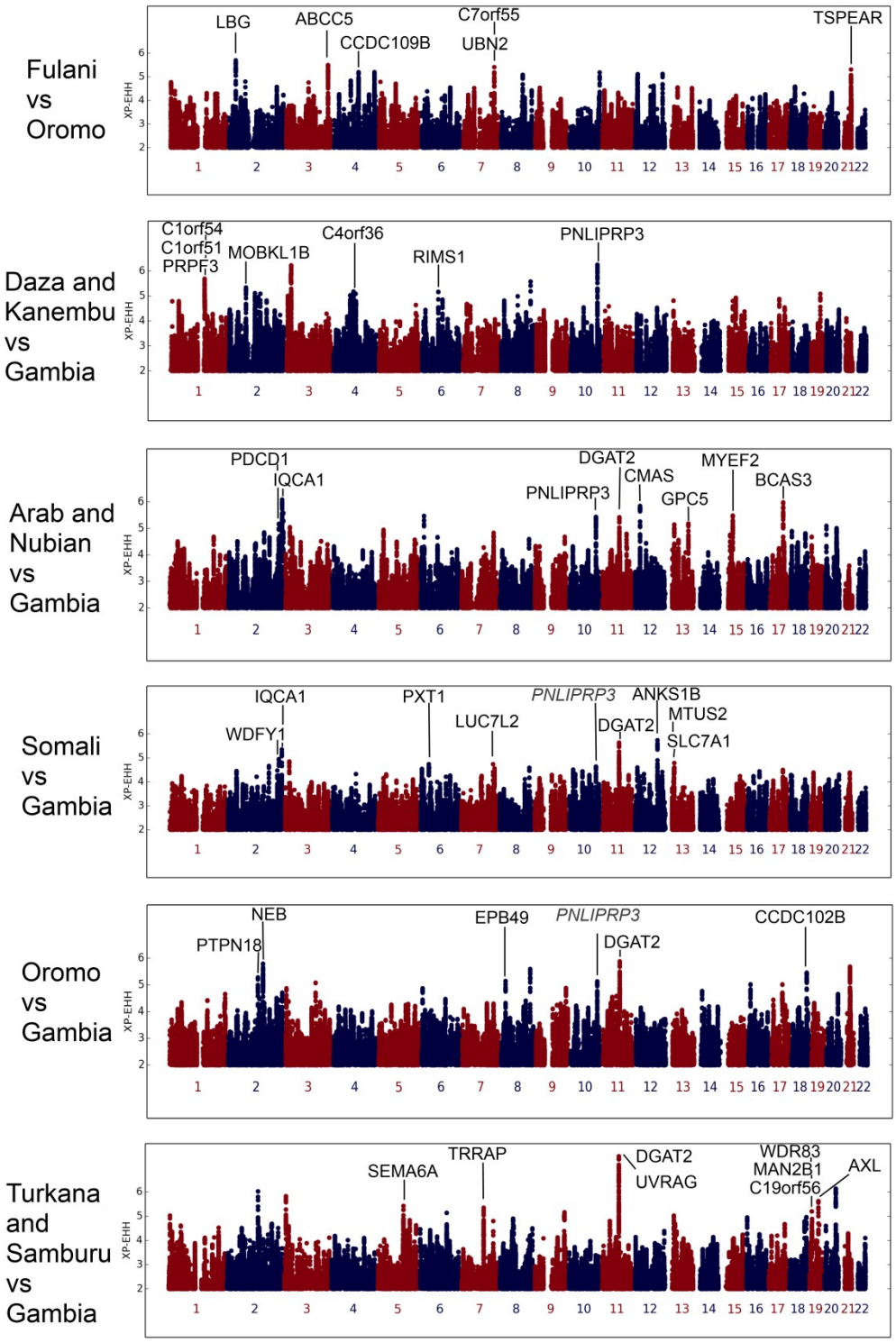

Figure S15 - Top-10 (in black letters) XP-EHH in Fulani vs Oromo, and Daza+Kanembu and each Eastern Sahelian population compared with the Western Gambia population. When some of the genes were also in the 0.1% significant tale of the distribution in other populations, although not in the top-10, they were represented in gray and italic letters.

Table S1 - Populations genotyped in this study. Sample size, country, subsistence system, language and geographic coordinates.

African Region	Population	Sample size	Country	Subsistence system	Language phylum	Longitude	Latitude
West Sahel	Gurmantche	15	Burkina Faso	Sedentarian	Niger-Congo	0.72	11.24
	Gurunsi	16	Burkina Faso	Sedentarian	Niger-Congo	-1.14	11.17
	Mossi	17	Burkina Faso	Sedentarian	Niger-Congo	-1.34	12.59
	Songhai	3	Mali	Sedentarian	Nilo-Saharan	-1.70	15.28
West-Central Sahel	Fulani	13	Burkina Faso; Niger; Chad	Nomadic	Niger-Congo	various	various
Central Sahel	Daza	18	Chad	Seminomadic	Nilo-Saharan	20.55	18.18
	Kanembu	5	Chad	Sedentarian	Nilo-Saharan	15.31	14.12
East Sahel	Arabs	16	Sudan	Sedentarian	Afro-Asiatic	30.75	18.41
	Nubians	14	Sudan	Sedentarian	Nilo-Saharan	30.48	20.79
	Oromo	17	Ethiopia	Sedentarian	Afro-Asiatic	35.05*	8.04*
	Somali	11	Somalia	Sedentarian	Afro-Asiatic	42.63*	2.00*
	Turkana	6	Kenya	Nomadic	Nilo-Saharan	36.71	2.74
	Samburu	10	Kenya	Nomadic	Nilo-Saharan	36.72	2.74
Total		161					

* These samples were collected in a refugee camp in Yemen; these coordinates reflect main geographical distributions of these populations.

Table S2 - Populations from other datasets used in this study. Sample size, country and reference.

Region	Population	Sample size	Country	Reference
Sub-Saharan Africa	Gambian	50	Burkina Faso	1000 Genomes Project
	Mende	50	Sierra Leone	1000 Genomes Project
	Yoruba	50	Nigeria	1000 Genomes Project
	Esan	50	Nigeria	1000 Genomes Project
	Luhya	50	Kenya	1000 Genomes Project
	Mandenka	22	Senegal	Li et al. (2008)
North Africa	Mozabite	27	Algeria	Li et al. (2008)
Near East	Bedouin	45	Israel	Li et al. (2008)
	Palestinian	46	Israel	Li et al. (2008)
Europe	Sardinian	28	Italy	Li et al. (2008)
	Tuscany	50	Italy	1000 Genomes Project
	French	28	France	Li et al. (2008)

7.2 Appendix B – Supplementary Material Paper II.

OSBPL10, RXRA and lipid metabolism confer African-ancestry protection against dengue haemorrhagic fever in admixed Cubans.

In preparation.

Due to the large extend of Supplementary Material, only the most relevant supplementary tables and figures are presented in printed version. Complete Supplementary Material is provided in digital format (Supplementary Tables – Paper II).

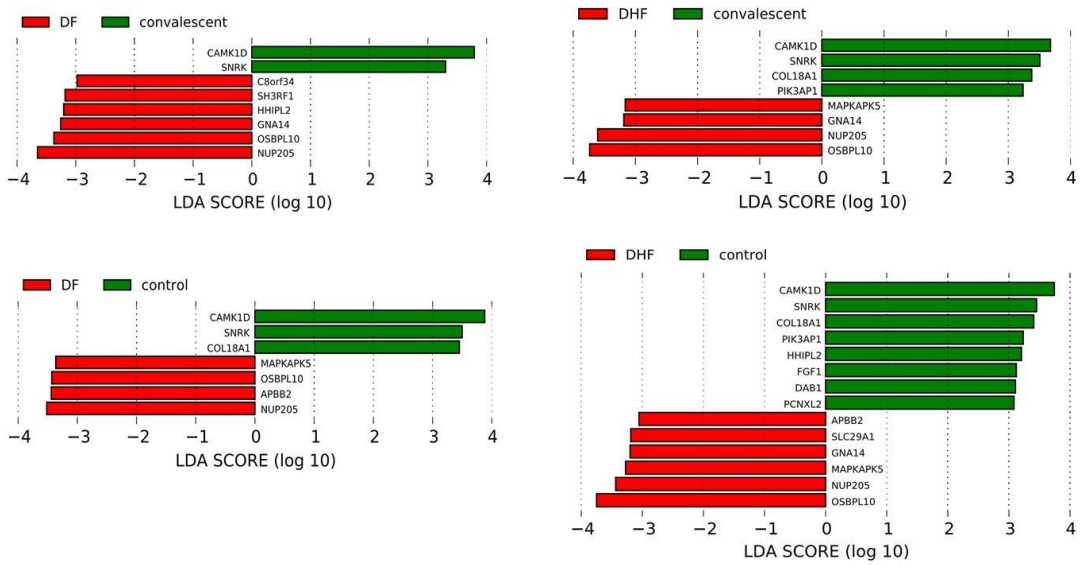

Figure S7 - LefSe for African-related associated genes with dengue fever in the three comparison groups (HCG, FCG and OCG).

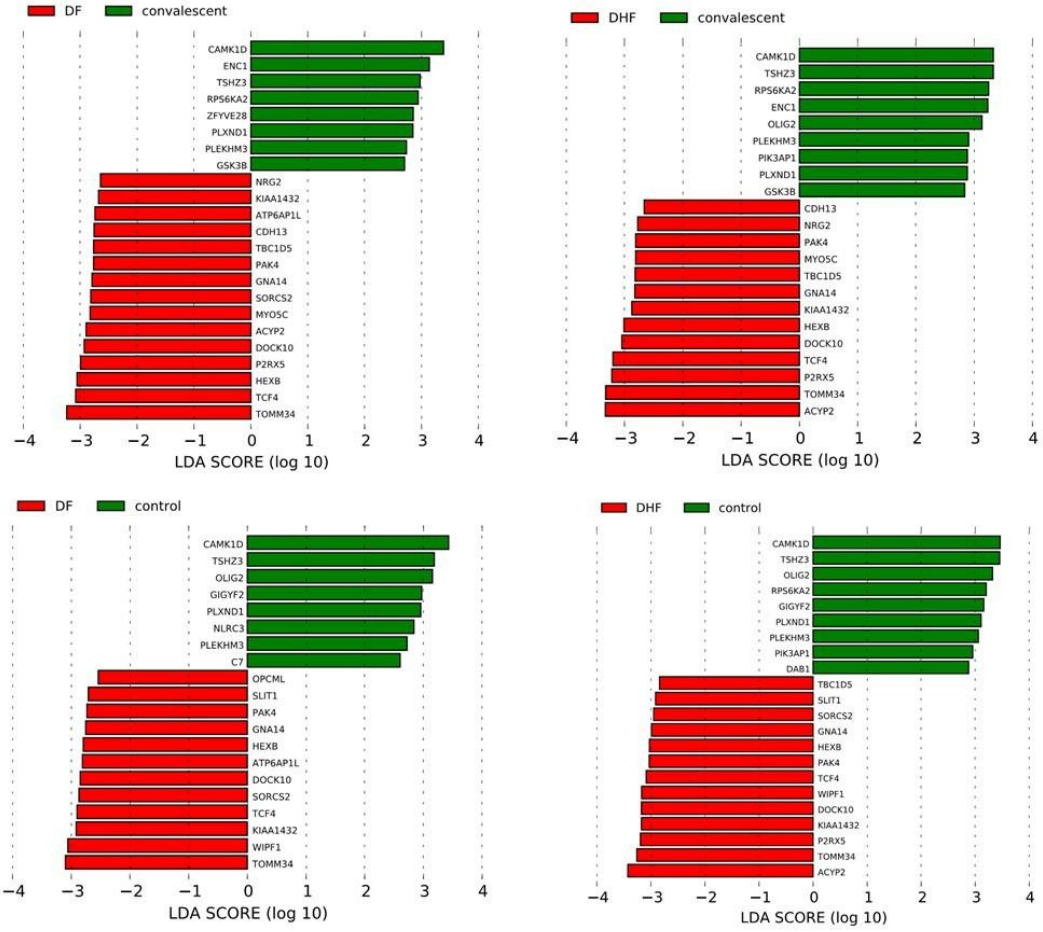

Figure S8 - LefSe for non-African-related associated genes with dengue fever in the three comparison groups (HCG, FCG and OCG).

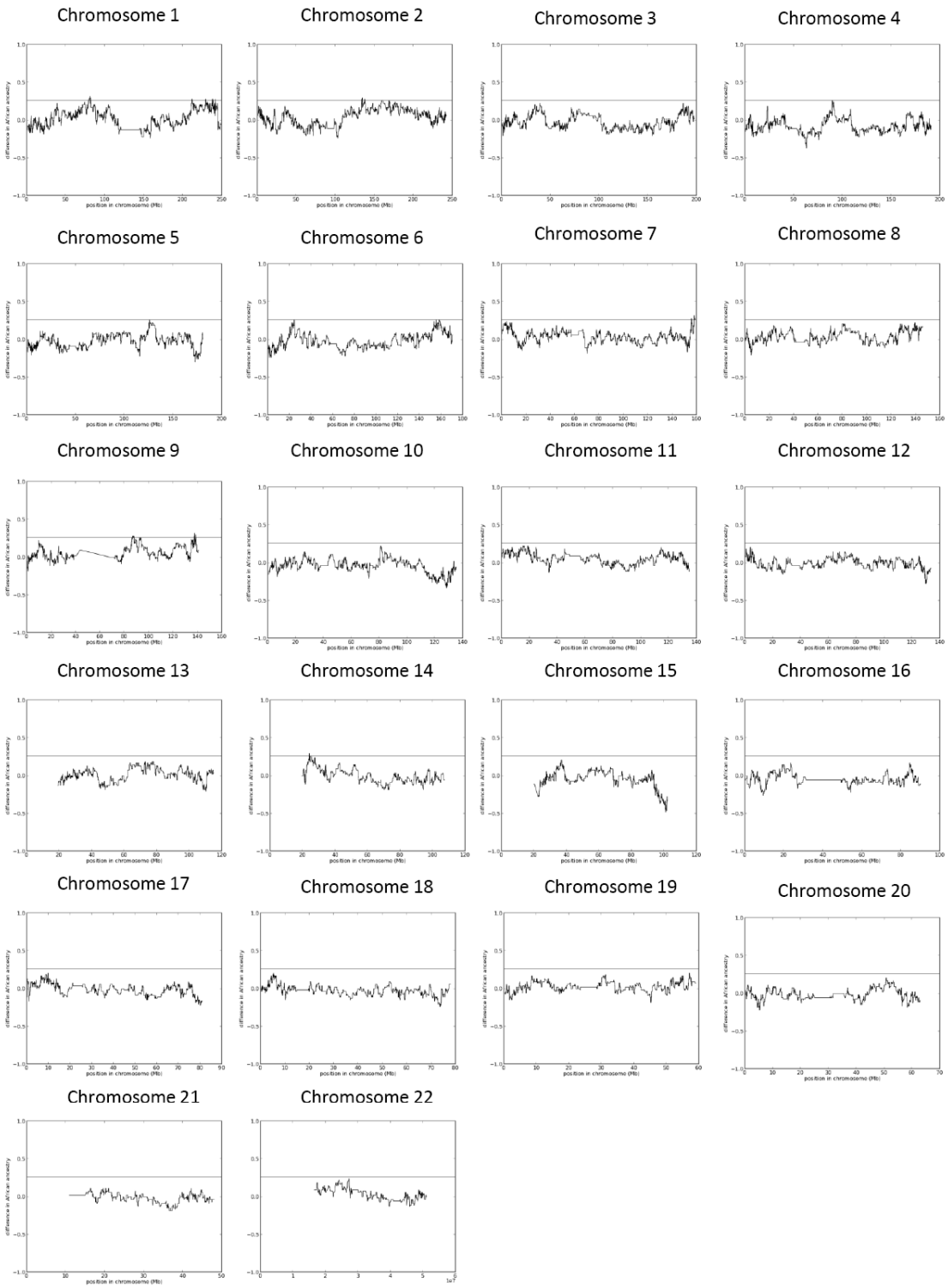

Figure S9 - Difference in African ancestry in HCG along the 22 autosomes. The line defines the 99% confidence interval.

Table S7 - Regions along chromosomes identified in RFMix analysis in the HCG above the 99% confidence interval for the difference in African ancestry.

Chr	Block start	Block end	No. of African blocks in asymptomatic control	No. of African blocks in hemorrhagic	p-value	Protein coding genes
1	79666332	80267726	33	18	0.024319984	
1	80269886	80916102	33	17	0.015016031	
1	80921812	81159163	34	17	0.009941148	
1	81159521	81352249	32	17	0.022331877	
1	81655249	81768121	33	18	0.024319984	
1	81768971	81892985	33	17	0.015016031	
1	81895097	81989420	31	16	0.020339182	
1	211798430	212093158	33	18	0.024319984	NEK2 LPGAT1
1	229499636	238391155	32	17	0.022331877	ACTA1 NUP133 ABCB10 TAF5L C1orf198 URB2 GALNT2 PGBD5 COG2 AGT CAPN9 TTC13 ARV1 FAM89A TSNAX DISC1 TRIM67 C1orf131 GNPAT EXOC8 SPRTN EGLN1 DISC2 SIPA1L2 MAP10 NTPCR PCNXL2 KIAA1804 KCNK1 AK054726 SLC35F3 COA6 TARBP1 IRF2BP2 TOMM20 RBM34 ARID4B GGPS1 TBCE B3GALNT2 AX747026 LYST GNG4 GPR137B NID1 ERO1LB EDARADD LGALS8 HEATR1 ACTN2 MTR MT1HL1 RYR2 ZP4
1	242010116	242077448	27	12	0.01257109	EXO1
2	134054312	134213436	28	13	0.014454423	NCKAP5
2	134363401	134589967	29	14	0.01638702	
2	134591340	134643199	30	14	0.010725139	
2	137591618	137769652	29	14	0.01638702	THSD7B
7	155921264	155967382	35	20	0.028242831	
7	158069264	158176618	36	19	0.012049956	PTPRN2
7	158178405	158536345	36	21	0.030160496	PTPRN2 NCAPG2 ESYT2
7	158763288	159124481	37	22	0.032039229	VIPR2
9	86676854	86771998	35	20	0.028242831	
9	86988321	87289280	34	19	0.02629332	NTRK2
9	87292328	87530935	35	20	0.028242831	NTRK2
9	87694716	87939277	34	19	0.02629332	
9	137516213	137694369	34	19	0.02629332	RXRA-COL5A1
9	137699002	137730956	34	17	0.009941148	COL5A1
9	137732458	137780823	33	18	0.024319984	FCN2 COL5A1
9	138668073	138817178	35	19	0.017938274	KCNT1 CAMSAP1
14	24114490	24159882	30	14	0.010725139	DHRS2

7.3 Appendix C – Supplementary Material Paper III.

Genetic Stratigraphy of Key Demographic Events in Arabia

PloS One, *10*(3), e0118625.

Due to the large extend of Supplementary Material, only the data related to genome-wide analysis are presented. Complete Supplementary Material is provided in digital format (Supplementary Tables – Paper III).

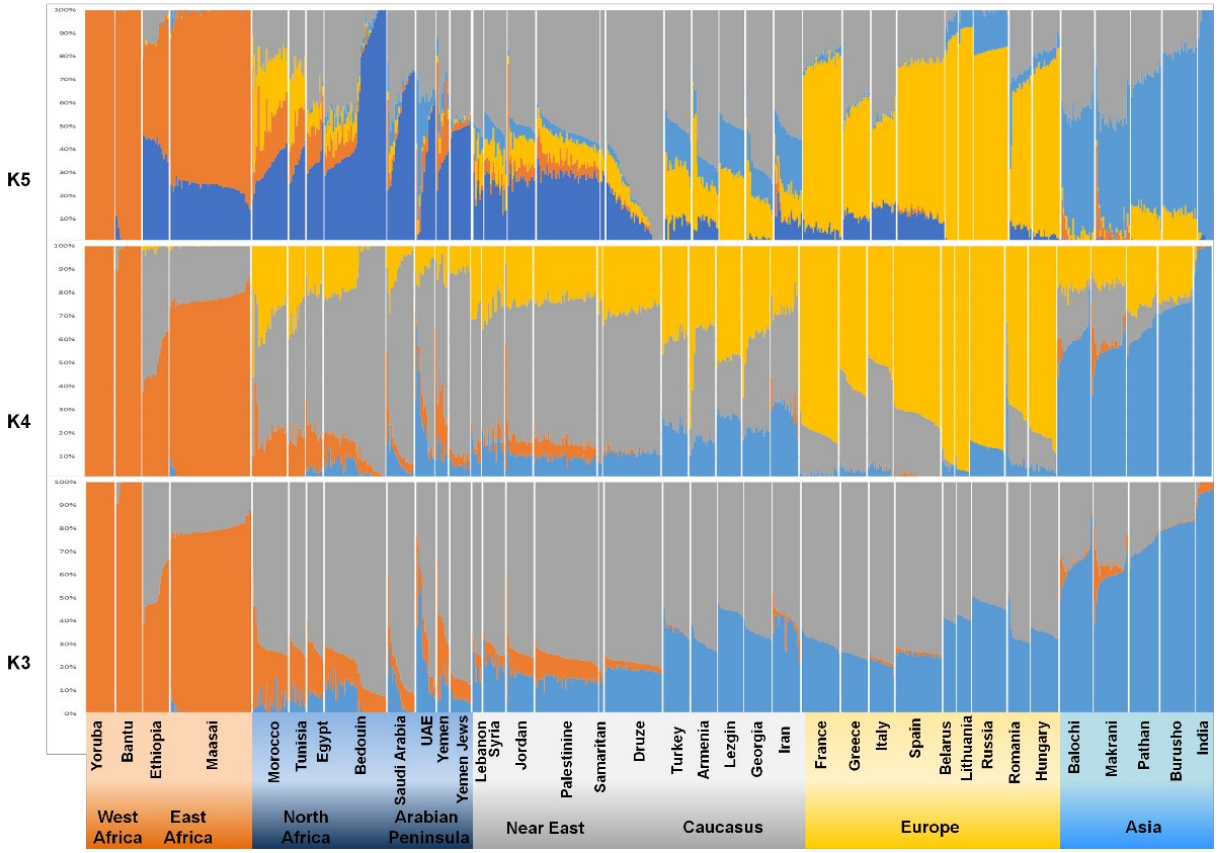

Figure S38 - Population structure inferred by ADMIXTURE analysis. Each individual is represented by a vertical (100%) stacked column of genetic components proportions shown in colour for K = 3, 4 and 5.